Time Series Methods for the Simulation of Wind Speed Fields Across Great Britain

Edwards, Gruffudd

Award date:
2014

Awarding institution:
University of Bath

Time Series Methods for the Simulation of Wind Speed Fields across Great Britain

Submitted by
Gruffudd A. Edwards, M.Sci., P.G.Dip.
for the degree of Doctor of Philosophy
of the University of Bath
Department of Electronic and Electrical Engineering
2013

Gruffudd A. Edwards

Bath, May 2013

Summary

Many modern power systems are undergoing a transformation involving the addition of large amounts of renewable energy generation capacity, particularly wind generation. Looking towards 2050 or further, it may be that the majority of society's energy needs will be met by them. The outputs of renewable energy generators are only partially predictable and cannot be increased to meet demand. Therefore, in order to maintain the high levels of power system reliability currently enjoyed (in developed countries, at least) future power systems will have to rely on several methods such as storage, international interconnectors, deferral of demand and highly flexible conventional plant.

The challenge partially addressed here is how to calculate, for some future electricity network scenario, whether the proposed combination of such approaches would be sufficient to maintain acceptable levels of reliability. It has recently been established in the literature that such assessments require sequential Monte Carlo simulation of the entire power system, including finite transmission capacities. Further, accurate representation of spatio-temporal patterns for the availabilities of the renewable generators is essential for such simulations. It is the goal of this PhD project to explore ways of accurately representing the spatio-temporal patterns of wind speeds in Great Britain (GB) accurately.

Since we wish to maintain the very high levels of reliability displayed by current power systems, occasions where available generation and transmission capacities cannot meet demands should remain very rare. However, we wish to associate precise probabilities to such events. Doing so means that the Monte Carlo simulations must be very long, typically involving 100,000 of the most 'risky' hours. For such long simulations, the availability of renewable generators may be represented either by historical weather records, repeated many times, or by a time series model that can keep on generating 'new' synthetic data indefinitely. Since weather may be conceived as a random process, it is the underlying rationale of this research project that it is best to capture the nature of the process generating the meteorological quantities of interest within a suitable model, and to use that model to generate new realisations of the process.

This thesis therefore details research concerned with the development of such a time series model, and associated algorithms, capable of generating synthetic wind speed datasets. Specifically, we are concerned with representing the wind energy resource available to generators connected to the Great Britain (GB) electricity networks, by means of an hourly averaged wind speed field across the Country. The model must capture patterns of variability in the resource occurring on many timescales – from hourly fluctuations to gradual climatic shifts. In order to be matched with demand and the availability of other resources, the synthetic data are time-stamped with time of the day and day of the year. All deterministic and stochastic patterns relating to specific times must therefore be accurately reproduced. The data used in this research are wind speed recordings taken by the UK's Met Office, and they are assumed to perfectly represent the nature of the wind resource at resolutions of 1 hour and above.

Complete accuracy in characterising and reproducing the resource's behaviours is impossible, particularly in a multivariate context with a large number of dimensions. Clearly there exists an optimum level of complexity that must be established. Previous wind speed modelling work conducted at the University of Bath was novel in that it represented simultaneous wind speeds at a large number (20) of locations across GB. However, the previous Bath model was concerned with the recourse during winter only and was not capable of associating specific time-stamps to the generated speeds. It is shown in this report that the wind resource displays complex stochastic behaviour in relation to short-term volatility clustering, a feature not adequately captured by the Bath model, nor indeed any previous multivariate model of wind speed. The same can be said about the long-term variability displayed by the resource, with changes occurring from month-to-month and year-to-year considered here, and proving extremely difficult to capture and reproduce.

The final model structure proposed here is the 2-factor-VGARMA-APARCH, along with various transformations and representation of deterministic seasonalities on annual and diurnal scales. This model structure is at the cutting-edge of time series modelling, and is shown to be reasonably good at reproducing all aspects described above, albeit with considerable scope for improvement in future work. Literature reviews reveal that the optimum level of complexity with regard to some aspect of wind speed dynamics is determined by the purpose of the model. For example, if the purpose is to calculate the 99% reliable baseload provided by a spatial arrangement of wind farms, then facing the formidable

challenge of fitting high-dimensional and heterogeneous pair-copula vines to describe the spatial structure of model residuals may be justified – but in general it is not.

A hierarchical model, probably involving 3 layers was a possible modelling alternative. The lowest level of this model would have been the wind speeds, with their stochastic dynamics determined by a model structure and a set of parameters. The 2nd level would have involved a latent variable representing the type of atmospheric circulation occurring (e.g. 'westerly', 'high pressure ridge'), and its value would determine the parameter set for the wind speed dynamics. A 3rd level would have involved an additional latent variable, dictating the dynamics of the first latent variable, and could possibly be necessary in order to account for slow climatic variability. Such a scheme is probably the only way of capturing the very low frequency dynamics of the wind speed field more accurately than the path followed by the project. Although this aspect of the wind speed field's behaviour is a priority for the current research, this approach was deemed too complex to be achievable in a single PhD project.

There is even evidence, discussed in Chapter 8, that a much simpler model – that avoids explicit modelling of low frequency variability and volatility clustering, would be superior in some respects. That is probably inevitable, particularly for a high-dimensional model, where a few statistical 'fudges' were necessary in order for parameter fitting to be feasible.

The thesis contains a presentation of relevant time series modelling theory and methods; literature reviews of both the nature of the wind resource and previous attempts at modelling it; further statistical analysis of the GB resource; presentation of the parameter fitting methodology and consequent model evolution; simulation methodology and analysis of performance. Despite the main purpose of the research being the generation of wind *speeds*, a wind *power* dataset is generated for an example wind capacity distribution scenario.

Acknowledgements

I would like to express much gratitude to my supervisor Dr R.W. Dunn for his guidance, and providing me with the opportunity to undertake this course of research. Thanks are also due to Professor Furong Li and Dr Miles Redfern for their feedback during the M.Phil. – PhD. transfer process.

This work has benefited from many conversations with fellow PhD candidates from the Supergen FlexNet consortium, most notably Dr A. Sturt, formerly of Imperial College. Advice was also gratefully received from academics within the consortium, particularly Professor G. Ault and Dr K. Bell from the University of Strathclyde and Professor G. Harrison from the University of Edinburgh.

Essential meteorological perspective was very generously provided by academics at the Department of Meteorology at the University of Reading, in particular Dr D. Brayshaw. Statistical advice was provided by several members of the Department of Actuarial Mathematics and Statistics at Heriot-Watt University, including Dr S. Zachary and Dr J. Cruise; the moral support of the former being greatly appreciated.

I am very grateful to friends and family for their continued support and encouragement, my parents in particular.

This work was funded by the EPSRC as part of the Supergen FlexNet consortium.

Contents

List of Figures

List of Tables

Abbreviations

ACF	Auto-Correlation Function
ACS	Average Cold Spell (demand level)
AIC	Akaike's Information Criterion
APARCH	Asymmetric Power Auto-Regressive Conditional Heteroskedasticity (process)
AR	Auto-Regressive (process)
ARCH	Auto-Regressive Conditional Heteroskedasticity (process)
ARFIMA	Auto-Regressive Fractionally Integrated Moving Average (process)
ARMA	Auto-Regressive Moving Average
ATLB	Atlantic Blocking (atmospheric circulation type)
ATLR	Atlantic Ridge (atmospheric circulation type)
ATLT	Atlantic Trough (atmospheric circulation type)
BADC	British Atmospheric Data Centre
BIC	Bayesian Information Criterion
BL-GARCH	Bi-Linear Generalised Auto-Regressive Conditional Heteroskedasticity (process)
BWM	Bath Wind Model
CCF	Cross-Correlation Function
CDF	Cumulative Distribution Function (probability)
CHP	Combined Heat and Power (Generation)
CO2	Carbon Dioxide
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
EFC	Effective Firm Capacity (of generator)
ELCC	Effective Load Carrying Capacity (of generator)
EOF	Empirical Orthogonal Function (in Principal Component Analysis)
EURB	European Blocking (atmospheric circulation type)
EURR	European Ridge (atmospheric circulation type)
EURT	European Trough (atmospheric circulation type)
FIAPARCH	Fractionally Integrated Asymmetric Power Auto-Regressive Conditional Heteroskedasticity (process)
GARCH	Generalised Auto-Regressive Conditional Heteroskedasticity (process)
GARMA	Generalised Auto-Regressive Moving Average (process)

GB	Great Britain
GED	Generalised Error Distribution
GH	Garrad Hassan & Partners Ltd
GW	Giga Watt
GWL	Grosswetterlagen (atmospheric circulation classification system)
HB-GWL	Hess and Brezowsky's Grosswetterlagen catalogue (of circulation)
HMM	Hidden Markov Model
IEEE	Institute of Electrical and Electronics Engineers
i.i.d	independent and identically distributed (process)
KDE	Kernel Density Estimates/ Estimation
LENS	Long-term Electricity Networks Scenarios (project)
l.h.s.	left hand side
LOEE	Loss of Energy Expectation
LOLE	Loss of Load Expectation
LOLP	Loss of Load Probability
MLE	Maximum Likelihood Estimation
MODWT	Maximum Overlap Discrete Wavelet Transform
MSLP	Mean Sea Level Pressure
MVN	Multivariate Normal
MW	Mega Watt
NAO	North Atlantic Oscillation (phenomenon)
NAOI	North Atlantic Oscillation Index
NG	National Grid plc
NREL	National Renewable Energy Laboratory (United States)
OLS	Ordinary Least Squares
PACF	Partial Auto-Correlation Function
PC	Principle Component (of a dataset)
PCA	Principal Component Analysis
PDF	Probability Density Function
QMLE	Quasi Maximum Likelihood Estimation
r.h.s.	right hand side
r.m.s	root mean square
SARIMA	Seasonal Auto-Regressive Integrated Moving Average (process)
SARMA	Seasonal Auto-Regressive Moving Average (process)

s.d.	standard deviation
SLP	Sea Level Pressure
SO	System Operator
TGARCH (process)	Threshold Generalised Auto-Regressive Conditional Heteroskedasticity
UK	United Kingdom
VAR	Vector Auto-Regressive (process)
VARMA	Vector Auto-Regressive Moving Average (process)
VCS	Very Cold Spell (demand level)
VTE	Variance Targeting Estimation
W	Westerly (atmospheric circulation type)

List of Symbols

$N(0,1)$	The standard normal distribution
U_t	The wind speed at time t in a regression model, used in chapter 1
$\underline{U_t}$	The vector of zonal wind speeds at time t in a regression model, used in chapter 1
$U_{i,t}$	The i^{th} component of a vector of zonal wind speeds at time t in a regression model, used in chapter 1
φ_i	The i^{th} auto regressive coefficient in an ARMA model
$\boldsymbol{\varphi}_i$	The i^{th} auto regressive coefficient matrix in a VARMA model
Z_t	The noise term at time t in an ARMA model: the random innovation in a simulation context and a residual in a model fitting context.
$\underline{Z_t}$	The vector of noise terms at time t in a VARMA model – interpretation as above
X_t	The value of a generic random process $\{X_t\}$ at time t
p_{ij}	The probability that in a Markov chain, the variable will make a transition from state i to state j as the time index progresses
i	A general integer index variable
i	The unit imaginary number
j	A general integer index variable
k	A general integer index variable
t	The index for time units
t	A continuous variable representing time
x	Generic continuous variable
y	Generic continuous variable, when x is not available
τ_n	The random residence time before the n^{th} transition in a semi-Markov process

M	A random number representing the excess in total available generation over total demand
$F_M(\cdot)$	The cumulative probability distribution function for M
$f_M(\cdot)$	The probability density function for M
Y	A random number representing the additional power availability offered by a generator
μ	The expectation value of a generic random process
$\mu_{t\|I}$	The conditional expectation of a generic random process at time t, given all past values
μ_Y	The expectation value of Y
σ^2	A general symbol for variance
σ_Y^2	The variance of Y
X	A random number representing the total power available from 'background' generators in a system of interest
$f_X(\cdot)$	The probability density function for X
D	A random number representing the total demand in a system of interest
$f_D(\cdot)$	The probability density function for D
k_W	The shape parameter of a Weibull probability distribution
C_W	The scale parameter of a Weibull probability distribution
$f_W(\cdot\,;\, k_W, C_W)$	The probability density function for a Weibull distributed random variable, with shape parameter k_W and scale parameter C_W
$Q_W(\cdot\,;\, k_W, C_W)$	The reverse cumulative probability distribution function for a Weibull distributed random variable, with shape parameter k_W and scale parameter C_W
δ_W	An exponent used to render a Weibull distributed random variable roughly Gaussian
z	Height above the ground
u	A general wind speed variable

u^*	Wind shear velocity
k_{VK}	Von Kármán's constant (used in wind-shear calculations)
z_0	The surface roughness length
ψ_S	An atmospheric stability correction term (used in wind-shear calculations)
τ	Atmospheric shear stress
ρ	Air density
α	Wind shear exponent
d_p	The distance between a pair of sites where wind speeds were recorded
d	Differencing order, or parameter for long memory processes. Also known as the 'long-memory parameter' in the latter case
$\gamma_X(k)$	The auto-covariance function for the random process $\{X_t\}$
$\rho_X(k)$	The auto-correlation function for the random process $\{X_t\}$
$\boldsymbol{\Gamma_X}(k)$	The auto-covariance matrix function for the random process $\{X_t\}$
$\boldsymbol{\rho_X}(k)$	The auto-correlation matrix function for the random process $\{X_t\}$
$\boldsymbol{V}$	A diagonal matrix of variances
ω	Angular frequency
ω_j	The j^{th} Fourier angular frequency
$f_{\omega,X}(\omega)$	The spectral density function for the random process $\{X_t\}$
$\boldsymbol{f}_{\omega,X}(\omega)$	The spectral density matrix function for the vactor random process $\{\underline{X}_t\}$
p	The maximum auto-regressive lag in an AR or ARMA model
q	The maximum lag of past-error regression in a MA or ARMA model
B	The backwards shift operator in a regression model
$\Phi(B)$	The AR polynomial for an ARMA model
$\theta(B)$	The MA polynomial for an ARMA model
σ_z^2	Variance of the noise Z for an ARMA model
$\boldsymbol{\Phi}(B)$	The AR polynomial for a VARMA model
$\boldsymbol{\Theta}(B)$	The MA polynomial for a VARMA model

$\boldsymbol{\Sigma}_Z$	The covariance matrix for the noise vector $\underline{Z}$ for a VARMA model
ψ_j	The regression coefficients, or equivalently filter weights, in an MA(∞) representation of an ARMA model
T	An integer representing the upper bound of a discrete time interval
Y_T	The sum of a random process from $t = 1$ to $t = T$
Y_t	The value of the general random process $\{Y_t\}$ at time t (when X_t is not available)
$T(\omega)$	A power transfer function connecting the spectra of two processes
D_s	Seasonal differencing order for SARIMA models
λ_{BC}	Box-Cox transformation coefficient
ω_G	A Gagenbauer frequency
v	The cosine of a Gagenbauer frequency
y	General continuous variable, when x is not available
δ	The long memory/ differencing parameter for a Gagenbuer process
δ_i	The long memory differencing parameter for zone i in a vector Gagenbauer process
$F_Y(y)$	The cumulative probability distribution function for a variable Y
$\underline{Y}$	A general random vector
Y_i	The i^{th} component of the vector $\underline{Y}$
n	An integer variable indicating the number of elements in a set
N	An integer representing the number of elements in a specific set
r_b	A positive continuous variable used in a generalised binomial expansion
$C_j(\delta, v)$	The j^{th} Gagenbauer polynomial
r	The order of the finite response filter in a GARCH model
s	The order of the infinite response filter in a GARCH model
α_0	A constant term in the definition of a GARCH process
α_i	The i^{th} regression coefficient of the finite response filter in a GARCH model

$\beta(B)$	The infinite response filter of a GARCH model in polynomial form
β_j	The j^{th} regression coefficient of the infinite response filter in a GARCH model
h_t	The conditional variance of the random innovation at time t in a GARCH model
ξ_t	Model innovations normalised by the conditional variance
$\underset{\sim}{\xi}_t$	The vector of independent random processes, all temporally i.i.d., which drive a heteroskedastic innovations process with multivariate conditionality
w	The order of the asymmetry term summation in BL-GARCH models
γ_k	The k^{th} coefficient in the asymmetry term summation in BL-GARCH models
α^+	The finite response filter coefficient for positive innovations in a TGARCH model
α^-	The infinite response filter coefficient for negative innovations in a TGARCH model
λ	The asymmetry parameter in an APARCH model
ι	The 'power' exponent in an APARCH model
$\boldsymbol{H}_t$	The innovations covariance matrix for a vector random process with multivariate conditionality
$\underline{h}_t$	The lower triangular portion of conditional covariance matrix $\boldsymbol{H}_t$ stacked into a column vector
$\underline{\eta}_t$	The lower triangular portion of the matrix formed by the product $\underline{z}_t . \underline{z}_t'$ stacked into a column vector
$\boldsymbol{A}$	A matrix of parameters used in the definition of VEC(1,1) multivariate hetersoskedastic models
$\boldsymbol{G}$	A matrix of parameters used in the definition of VEC(1,1) multivariate hetersoskedastic models
$\underline{c}$	A vector of parameters used in the definition of VEC(1,1) multivariate hetersoskedastic models
$\widehat{\Upsilon_X}(k)$	An estimator for $\Upsilon_X(k)$ based on a finite sample

$\underline{\widehat{\Phi}_p}$	A vector of regression parameter estimates for an AR(p) model
$\boldsymbol{\Gamma_p}$	A p x p matrix of auto-correlations, with lags ranging from $1-p$ to $p-1$
$\underline{\Gamma_{(p)}}$	A vector of auto-correlations, lags from 1 to p
$\hat{\sigma}_z{}^2$	An estimator of the innovation variance for a model, based on a finite sample
$\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$	An estimator of the innovation covariance matrix for a multivariate model, based on a finite sample
$\widehat{\sigma}_m^2$	The noise variance estimate for an AR(m) model, fitted with the Levinson-Durbin algorithm
$\widehat{\varphi}_{m,j}$	The j^{th} regressive coefficient estimate orf an AR(m)model, fitted with the Levinson-Durbin algorithm
$\boldsymbol{A}_{p,k}$	The k^{th} regression matrix in a VAR(p) model, fitted with the multivariate Levinson-Durbin algorithm
$\boldsymbol{\Delta_p}$	A matrix constructed as part of the multivariate Levinson-Durbin algorithm
$\boldsymbol{V_p}$	A matrix constructed as part of the multivariate Levinson-Durbin algorithm
$\underline{x}_n$	The n-dimensional vector formed by a set of n sequential observations of a quantity, in the context of maximum likelihood estimation
$\underline{X}_n$	A n-dimensional random vector, of which $\underline{x}_n$ is one realisation
φ	A short-hand expression for the set of auto-regression coefficients in an ARMA model
θ	A short-hand expression for the set of moving average coefficients in an ARMA model
r_n	A convenient construct used in the derivation of a likelihood function
$\boldsymbol{V}_{j-1}$	The covariance matrix of the j^{th} vector of prediction errors in a multivariate sample of observations, used in a likelihood calculation context
$\hat{\rho}_z(k)$	An auto-correlation estimator for the model residuals from a sample
Q_h	The portmanteau statistic used in model residual testing
h	The number of auto-correlation estimates used in a portmanteau test on model residuals

α_C	$1 - \alpha_C$ is the confidence level in a statistical test
$x(t)$	A continuous function on the interval [0,1]
b_0	A constant term in a Fourier series expansion
b_k	The k^{th} cosine term coefficient in a Fourier series expansion
a_k	The k^{th} sine term coefficient in a Fourier series expansion
c_{jk}	A coefficient with index values j and k in a wavelet expansion
$\psi(t)$	A mother wavelet, the basis function in a wavelet expansion
$\Psi(\omega)$	The Fourier transform of the mother wavelet $\psi(t)$
u_{WT}	A continuous variable representing time shift in a wavelet transform
s_{WT}	A continuous variable representing time scale in a wavelet transform
$W(u_{WT}, s_{WT})$	A continuous wavelet transform
$\psi_{u,s}(t)$	A wavelet function with parameters u_{WT} and s_{WT}, and argument t
u_i	The value of the marginal CDF of X_i when it takes the value x_i, forming arguments of a cupula function
$C_{\underline{X}}(u_1, \ldots, u_m)$	An m-variate copula function
$c_{\underline{X}}(u_1, \ldots, u_m)$	An m-variate copula density function
s_M	The number of states in which the variable of interest may exist in a discrete Markov process
$\underline{\Pi_t}$	For a discrete Markov process with s_M states, this is a vector with s_M dimensions, with dimension j containing the probability that the variable of interest is in state j at time t
$\boldsymbol{Q}$	The transition matrix for a (univariate, 1st order) discrete Markov process
$\boldsymbol{Q}_i$	The i^{th} transition matrix for a discrete Markov process of order > 1
l	The order of a discrete Markov process
v_i	Weighting coefficients for an univariate discrete Markov process of order > 1
$v_{i,j}$	Weighting factors for a 1st order multivariate discrete Markov process

$\underline{\Pi_t}^{(j)}$	The state probability vector $\underline{\Pi_t}$ for the j^{th} variable in a multivariate Markov process
$\boldsymbol{Q}^{(j,k)}$	A transition probability matrix for a multivariate discrete Markov process, relating states in the k^{th} series to the states in the j^{th} series
$m(t)$	A deterministic function capturing seasonalities in the mean of a time series
$\sigma(t)$	A deterministic function capturing seasonalities in the variance of a time series
X_t^{Stat}	A time series considered stationary following removal of deterministic seasonalities in mean and variance from observations
X_t^{Real}	The real component of a Gaussian complex random variable
X_t^{Imag}	The imaginary component of a Gaussian complex random variable
$\boldsymbol{\Sigma_{HR}}$	A correlation matrix calculated as part of fitting a model to wind speeds in Ireland
$\boldsymbol{C_{\Sigma}}$	A matrix that renders spatially correlated series independent, used to fit a model to wind speeds in Ireland
$\Psi_F(B)$	The finite impulse response filter in a FIGARCH model of variance
d_{var}	The long memory parameter in a FIGARCH model of variance
χ_{CM}	The coefficient connecting the conditional mean and variance processes for ARCH-in-mean processes
Ψ	Shorthand for the entire set of parameters defining a time series model
$Q_{ML}(X,\Psi)$	Beran's quasi-maximum-likelihood estimator
η	A parameter list for a time series model, lacking the noise variance
$\widetilde{Q_{ML}}(X,\eta)$	Beran's further simplified quasi-maximum-likelihood estimator
v_t	A change of variable useful in the fitting of GARCH type models, page 212
κ_0	A parameter used in the fitting of GARCH(1,1) models through VTE
γ_0	The unconditional variance of a GARCH process, used in model fitting through VTE
χ	A variable introduced to simplify notation

$z(i,t)$	Alternative notation for the residual/ noise innovation at zone i at time t
$z_{flat}(i,t)$	Model residuals/ noise innovations following transformation to remove all deterministic seasonality
$m_d(i,t)$	The mean value of residual/ noise variance at day or hour of the year t, and zone i
$m_{norm}(i,t)$	A version of $m_d(i,t)$ normalised by the long-term sample value
$s_d(i,t)$	The inter-annual variability of residual/ noise variance at day or hour of the year t, and zone i
$r_d(i,t)$	The ratio of variability to mean in the residual/ noise variance at day or hour of the year t, and zone i
$r_{norm}(i,t)$	A version of $r_d(i,t)$ normalised by the long-term sample value
$r_{trans}(i,t)$	A power-transformed version of $r_{norm}(i,t)$
$a(i)$	The zonal exponent in the power transformation of $r_{norm}(i,t)$ to $r_{trans}(i,t)$
h_B	Kernel bandwidth
σ	Sample standard deviation
v	A shape parameter in Student-t and GED distributions
λ_v	A function of the shape parameter v, used in the characterisation of the PDF of GED distributed variables
Y_t	A random variable constructed to fix problems associated with the simulation of noise with volatility clustering
ν_t	Values taken by the random variable Y_t
κ_t	A necessary transformation of ν_t
z_max_i	The maximum absolute value for the model residuals found in zone i in the historical sample
$kde\_max\_h_i$	The maximum of the KDE for the i^{th} historical series
$kde\_max\_s_i$	The maximum of the KDE for the i^{th} synthetic series
$U_{i,t}$	The doubly-differenced transformed wind speed at time t and zone i, used in chapters 7 and 8

$V_{i,t}$	The single-differenced transformed wind speed at time t and zone i, used in chapters 7 and 8
ι_i	Zone-specific constants used in a power transformation based algorithm to correct the kurtosis of synthetic series

Chapter 1

Introduction

1.1. Research Objective

This research project is concerned with the development of models, and associated algorithms, capable of generating vector time series of wind-speeds. The generated series will be treated as representative of the entire wind-speed field across Great Britain (GB) during hourly periods, thus capturing the wind energy resource available to generators connected to GB electricity networks. The country will be divided into a series of zones and each vector dimension will represent the wind speed at a location within a zone. The wind speed at any location within GB may then be extrapolated from its associated zonal value. More dimensions mean greater resolution for the field, but also imply a greater modelling challenge. Wind resource availability is assumed here to be stochastic in nature and the algorithms developed will produce synthetic datasets of unlimited size using purely statistical time series models.

The goal is to carefully select the model structure to ensure that the resources' chronological characteristics and spatiotemporal patterns, as found in historical data, are accurately reproduced in the datasets. Variability in the resource occurring on all timescales – from turbulent fluctuations to climatic changes between decades will be represented. Each vector will have an associated time stamp with day of year and hour of day, and care will be taken to ensure that all observed patterns relating to these are accurately reflected in the conditional distributions among the generated data. This will make the datasets suitable for use in sequential Monte Carlo simulations of the GB electricity system – either the present system or future scenarios, which may include full consideration of network constraints.

Any statistical model is only capable of reproducing certain aspects of an extremely complex natural phenomenon, such as surface wind speed fields, to a high level of accuracy. In principle, more complex model structures can simultaneously capture a greater number of aspects, but only if parameter estimation is successful. Estimation can be very challenging, particularly in a multivariate context with a large number of dimensions. The main task of this

research project is therefore to identify, fit and simulate from a model with an optimum level of complexity in its structure. If the model is too simple, it will fail with regard to many aspects of a thorough validation. If it is too complex, parameter estimation can fail, of the associated computational expense might not be justified.

While the focus of this project is on generating accurate wind *speed* datasets for a carefully chosen set of locations, a large GB wind *power* production time series will also be generated for an example generation scenario. Novelty lies largely in the choice of model structure, derived through the application of advanced time series modelling techniques, usually reserved for econometrics. The fact that power values are stamped with time of the day and day of the year also represents a significant novelty.

This project is not concerned with the simulation of wind directions, nor does it consider the use of wind direction data in order to develop better models for speeds. This is because the introduction of wind directions would not only double the number of variables but probably introduce many highly nonlinear sets of relationships. If developing time series models for short-term forecasting purposes, the inclusion of wind direction would be a natural choice – but a fairly easy one since the direction data are essentially exogenous to the process being modelled. However, in a simulation context one must model relationships between speeds and directions at the same zones, directions at different zones, and speeds and directions at different zones – i.e. the problem quickly becomes intractable.

Unfortunately, the omission of wind directions implies a serious lack of interpretation of 'what's going on' meteorologically at some point in the synthetic series. This surely places serious limitations on the synthetic series' value in the context of Monte Carlo simulations involving demand levels. The extent of such limitations is an issue that will receive significant attention in this thesis.

1.2. Context – The Wind Industry and Changing GB Electricity Networks

1.2.1. The Growth of Wind Generation

The UK has an excellent wind resource, indeed the best in Europe [1]. This fact, combined with internationally coordinated efforts to reduce CO2 emissions and increased concern for long-term energy security, means that installed wind generation capacity is growing rapidly in the UK – and world-wide. At the time of writing (19/07/2012) the UK operational capacity is 6.84 GW, with 3.972 GW in construction, 6.548 GW given consent and a further 10.805 GW under planning consideration [2]. This makes wind an interesting and highly relevant area for research. The research reported her was concerned with GB only – i.e. Northern Ireland excluded, since it not synchronised with the grid supplying England, Scotland and Wales.

The UK Government is legally bound by European Union targets to generate 15% of all energy consumed in the UK from renewable sources by 2020 [3]. This is equivalent to an almost seven-fold increase in the share of renewables in energy generation in close to a decade. This certainly requires radical change and the means by which this target will be met, if successful, cannot be entirely predicted. However, the lead scenario in the UK Government's 2009 Renewable Energy Strategy [3] stipulates that more than 30% of our electricity will be generated from renewables, with more than two-thirds coming from onshore and offshore wind. A Strategic Environmental Assessment conducted for the strategy concluded that 25 GW of offshore development would be permissible, in addition to 8 GW in existing plans. To place this in context, the peak demand for electricity in GB lies at around 60 GW, and installed generation capacity is about 90 GW [4].

In the longer term, UK Government policy is even more radical. The 2008 Climate Change Act created legally binding 'carbon budgets' aiming to cut UK CO2 emissions by at least 80% by 2050. The Government's Low Carbon Transition Plan [5], states that it will be around 2030 that the UK sees "a step change in intermittent generation delivered through both large and small scale renewable plants". It is considered likely that the electricity network would have to be entirely decarbonised by this date in order to achieve the 80% goal by 2050.

The potential wind capacity is very substantial: the Crown Estate, responsible for managing substantial areas of land including the seabed surrounding GB, states that its planned land leasing rounds provide the potential to deliver around 47 GW of offshore wind

generation, with the 3rd round "marking the start of the biggest single programme in the world" [6]. The growth in offshore wind capacity seen already means that the UK may be considered a global leader in the field of marine energy.

1.2.2. The Changing GB Electricity System

The addition of substantial wind generation capacity will have a very large impact on the GB electricity system. On long timescales of years and decades, referred to in a power systems context as planning timescales, the location of the best resource in generally remote areas means that networks will require substantial reinforcement and extension, at a cost measured in tens of billions of £. On shorter timescales of hours up to a few days, referred to as operational timescales, the wind's variability will have a radical impact on the way the way dispatchable generators are operated, and ultimately which types of generators remain economically viable. The inherent unpredictability of wind will have impacts on real time operation of the system, with an increased need for reserve, partially realised through demand control and eventually storage. The variability is also likely to lead to increased interconnection with other European countries.

In addition to increased wind capacity there are many factors which will contribute to the transformation of the electricity system:

- A likely increase in the penetrations of other renewables such as biomass, hydro, solar, wave and tidal generators [3].
- The increased electrification of heating. An increased use of heat pumps and storage heaters could create additional demand, for example, while micro combined heat and power (CHP) units, possibly burning biomass, could reduce winter peaks, so this is a source of considerable uncertainty [7].
- The electrification of transport – increasing overall demand, but assisting the System Operator as a source of very short term storage for system frequency control.
- Ageing infrastructure - a large proportion of both generators and transmission/ distribution infrastructure are in need of replacement due to age.
- The development of 'smarter' grids – currently a major 'buzz' word. Enabled by the use of increasingly sophisticated ICT systems, a smarter grid is one which "Gives a better understanding of variations in power generation and demand, and allows us to use that information in a dynamic and interactive way to get more out of the system" [8]. In the relatively near future, this essentially means consumers becoming

somewhat price responsive in real time thanks to information conveyed by their smart meters.

So, there is little doubt that the system will undergo radical transformation of nearly all its aspects over the coming decades, and that there exists considerable uncertainty about the nature of those changes. Whatever the outcome may be, the ability to make well-informed decisions requires an increased understanding of the spatio-temporal behaviour of the wind resource on many scales.

1.2.3. Public Perceptions of Wind Energy

It is far from certain that the very large wind capacities proposed by policy makers will actually be built. Despite generally high levels of support for renewable energy among the public in the UK, there is a significant and growing proportion of people reacting very strongly and publicly to the rate at which installed wind capacity is increasing. Naturally, many objectors are residents of rural areas where the installation of large wind-farms or turbines has been proposed, and their perceived quality of life will surely be impacted. Others believe that serving a significant proportion of GB's energy demand through wind turbines must be a phantasy created to support a radical environmentalist agenda.

Such views are supported by erroneous media reports that wind energy does nothing to reduce CO2 emissions. A typical and common example is a phrase such as: "wind turbines only work at maximum capacity for a third of time, meaning they have to be backed up by other technologies such as coal and nuclear [13]". In February 2012, a group of more than 100 MPs from the UK's Conservative Party wrote public letter to Prime Minister David Cameron calling for a dramatic cut in subsidies to wind farms on the basis that it is "unwise to make consumers pay, through taxpayer subsidy, for inefficient and intermittent energy production that typifies onshore wind turbines" [14].

There also exists doubt among the UK public regarding the accuracy with which policy makers and academic researchers present facts about the nature of the wind resource. For example, a study by Stuart Young Consulting commissioned by the conservation charity *John Muir Trust* found that there is "a growing concern that wind generation may not be able to deliver the contribution which has been predicted and used in government projection assumptions. Several studies have indicated that output is often less than anticipated or was claimed, at the planning stage of development" [15]. The study looked at a relatively short period for which the resource was poorer than its long term average, and only used data from

Scotland, yet Young concludes that “It is clear from this analysis that wind cannot be relied upon to provide any significant level of generation at any defined time in the future”.

It therefore seems reasonable to adopt the view that there may be a purely symbolic element to the Government ambitions described above, particularly those referring to 2050. Political scientist Ingolfur Blühdorn goes further, characterising the apparently serious nature of the Government’s commitment to sustainability related issues as a ‘performance of seriousness’ essential to the function of a consumerist society. In his opinion, “despite their vociferous critique of merely symbolic politics and their declaratory resolve to take effective action, late-modern societies have neither the will nor the ability to *get serious*.” [16]. Whether this applies to wind generation will be revealed in the decades to come.

1.3. Research Project Context

1.3.1. The *SuperGen FlexNet* Consortium

This research project is part of the *SuperGen FlexNet* Consortium. *SuperGen* is an academic initiative (strapline: Sustainable Power Generation and Supply), with the aim of helping the UK to meet its environmental emissions targets through a radical improvement in the sustainability of power generation and supply. Multidisciplinary researchers from several Universities work in a range of consortia, each focused on specific programmes of work.

The FlexNet Consortium ran from October 2007 to March 2012 and was concerned with researching the future form of GB electricity networks. Much of the networks are now due for replacement and due to the enormous costs involved, and the long life of assets, it is vital that plans for replacements prove to be appropriate. The key is to develop, plan and build networks that are flexible enough to meet several divergent scenarios. The goal of FlexNet was to lay out the major steps that will lead to such flexible networks – encompassing issues ranging from the technical to economics, market design and public perception. Its strap-line is 'Thinking Networks', reflecting an intention to both think about networks and to develop networks that can 'think' for themselves [17]. The rationale for the consortium's work was consolidated by the developments in the UK Government's energy policy described above, which occurred after work began.

Supergen FlexNet consisted of researchers from eleven UK universities: Bath, Birmingham, Cambridge, Cardiff, Durham, Edinburgh, Exeter, Imperial College, Manchester, Strathclyde and Surrey; and its industrial partners: EDF Energy, National Grid plc (NG), CE Electric and Central Networks). The research programme was divided into eight work-streams and this research project is part of a work-stream named 'Shape and Size of Future Electricity Networks'. This work-stream examined the factors that will dictate the future form of the network and the degree of flexibility required. It was initially conceived as the starting point of the research programme, driving research in other work-streams that looked at how the flexibility may be delivered.

One research project within the work-stream was concerned with characterisation of GB's non-wind marine resources [18], while another was concerned with wind resource modelling [19] – both based at the University of Edinburgh. In the latter, meso-scale meteorological modelling techniques were used to produce detailed time series of both onshore and offshore wind velocities, essentially interpolating spatially between historical

wind series for a 10 year period. The deliverables from the Edinburgh and current research projects are highly complementary – the Edinburgh data providing more detailed spatial information, while the datasets generated here are more focussed on temporal behaviours.

Work-stream leader Dr Graham Ault took a leading role in the LENS Project, an Ofgem (UK's energy regulator) commissioned project to develop scenarios specifically for the GB electricity networks in 2050 [21]. According to the FlexNet project website [17]: "The work performed by this [Future Shape and Size] work-stream has allowed us to support Ofgem in delivering the 'Long-term Electricity Networks Scenarios (LENS)' project. The tasks and deliverables over the first two years of FlexNet were fully aligned with the LENS project and the project team... worked closely with Ofgem over a period of approximately 18 months to March 2009."

The LENS scenarios represent significant novelty by virtue of being focussed specifically on future *networks*. However, they contain no more spatial detail than predecessors, and quantitative elements are derived entirely from economic modelling, which does not include spatiotemporal considerations. The establishment of a case for an 80% reduction in the UK CO2 emissions by 2050, as described above, represented a significant 'upping of the ante' for long term objectives. Even the most radical of scenarios with which the FlexNet Consortium has been engaged fall considerably short of such targets, but this by no means decreases the validity of the consortium's work.

Long datasets produced by the algorithms developed in this project are ideally suited to Monte-Carlo simulations that match renewable generation with demand, on a regional basis, and for plausible future scenarios. Examining the flows of energy implied could act as a form of validation, or the opposite, for scenarios such as those developed by LENS. However, full simulation of an electricity system, even under the steady-state conditions associated with hourly averages, is well beyond the scope of this research project.

1.3.2. The Bath Wind Model

A good starting point for researching the best model structure to be adopted in this project is the Bath Wind Model. The model is the result of work carried out by power systems researchers at the University of Bath, mainly during 2005 – 2006, notably Dr. Rod Dunn, Dr. Marcos Miranda and Professor Furong Li. The work formed part of FutureNet – a predecessor consortium to FlexNet within SuperGen [25].

The model methodology, as found in [25], involves representing the GB wind resource at hour *t* as a vector of zone-based average wind speeds $\underline{U}_t = [U_{1,t}, U_{2,t}, \ldots, U_{20,t}]^T$. Specific locations within the zones experience different wind speeds, but these values are assumed to have a fixed relationship with the representative zonal values. The methodology assumes that the set of wind speeds constitute a 4th-order vector autoregressive stochastic process – concisely written as a VAR(4) process. Historical records are single realisations of this process and future wind speeds are another – with the same parameters and with historical data providing starting values. In forming the vector $\underline{U}_t$ all component wind speeds had their mean values removed. So, we have a zero mean process satisfying the equation

$$\underline{U}_t = \boldsymbol{\varphi_1}\underline{U}_{t-1} + \boldsymbol{\varphi_2}\underline{U}_{t-2} + \boldsymbol{\varphi_3}\underline{U}_{t-3} + \boldsymbol{\varphi_4}\underline{U}_{t-4} + \underline{Z}_t \quad (1.1)$$

where the $\boldsymbol{\varphi_i}$ are autoregression coefficient matrices which correlate the values of each dimension in $\underline{U}_t$ to its own past values and also the other dimensions' past values; while $\underline{Z}_t$ is a vector of independent and identically distributed (i.i.d.) Gaussian white noise terms. The authors do not state whether they assume that the noise vector has a diagonal covariance matrix, but allowing correlation certainly adds flexibility to the modelling process.

The methodology divides GB into 20 zones, referred to here as the Bath Zones, which are based on 17 study zones adopted by NG, as presented in their Seven Year Statement [26], with some additions. The zones represent areas of the country with strong internal electrical connections, but with weaker interconnections to the rest of the system. A study involving the Bath Wind Model will generally be concerned with the balance of total generation and demand within zones and the resulting flows between them. Certain flows across the boundaries between zones, or group of zones, are of particular interest – for example the boundary along the Scottish border, separating the set of zones in Scotland from those in England and Wales. The 3 additional Bath Zones are in the North and North West of Scotland – the more detailed division of Scotland reflects the fact that it is, and will probably continue to be, home to a very significant proportion of wind generation capacity. In order to convert wind speeds into power outputs, speeds are first increased to account for both their location and the fact that turbines are taller than meteorological anemometers, before transformation into power outputs via a standard turbine power curve – as described in greater detail in later chapters. The Bath Zones are shown in Figure 1.1 below, in an image taken from a study commissioned by National Grid [27].

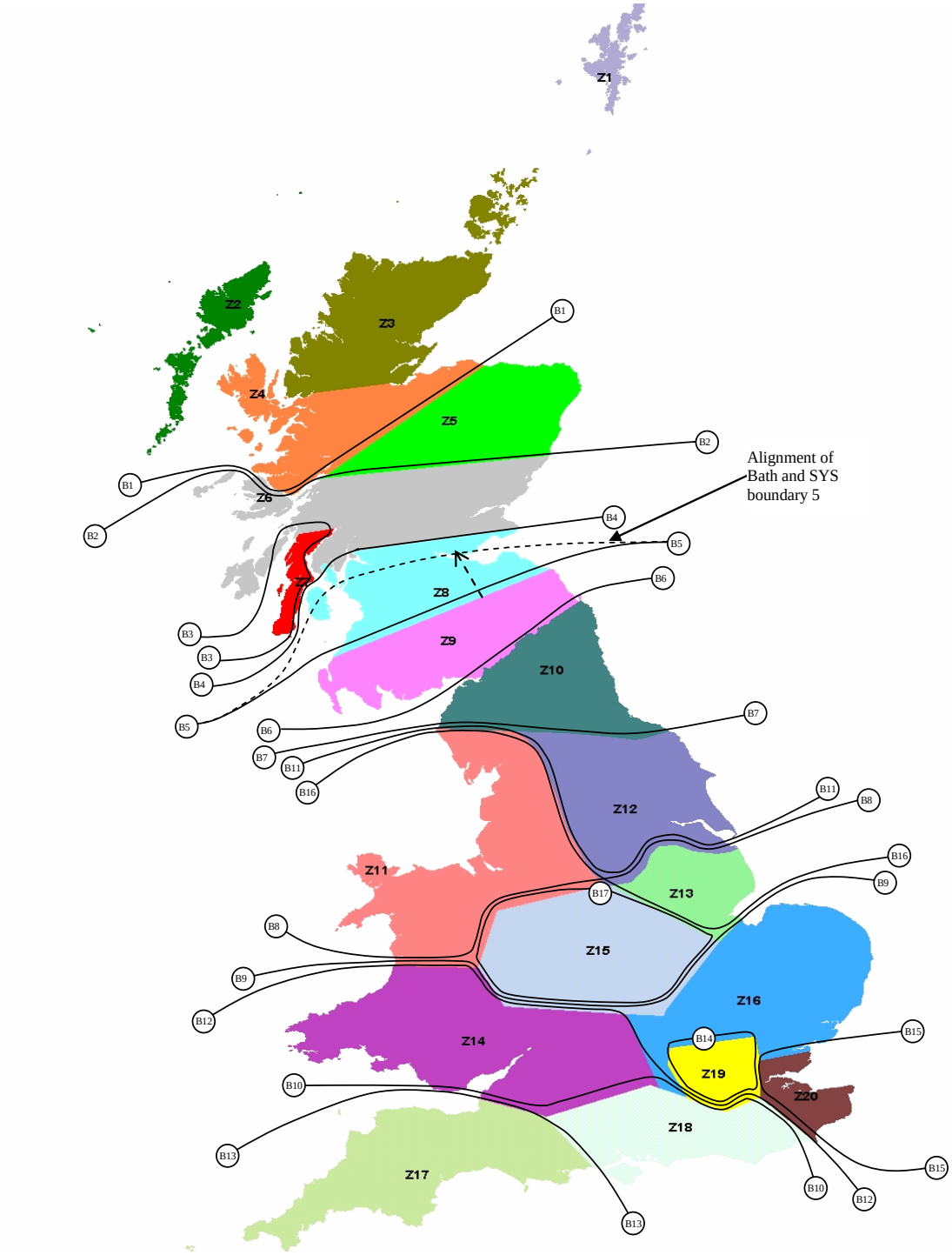

Figure 1.1. GB divided according to the Bath Zones.

1.4. Time Series Model Types

The Bath Model is an example of a regression-based time series model, in which values of the series, or set of series, are regressed upon their past values. Being stochastic in nature, such models may be conceived as temporal filters which may be applied to a process of i.i.d. random 'innovations' in order to produce the series of interest. The structure and parameter values of the filter, along with the choice of probability distribution for the innovations, ensure that the generated series display all the required temporal characteristics – or spatiotemporal characteristics in the case of vector series. This may be initially verified visually, and then via statistical tools such as spectra and correlation functions, as described in Chapter 3

In the multivariate case, if the collection of series of interest can be well approximated as having a multivariate normal (MVN) distribution, extension from single location to vector models is conceptually quite straightforward, although parameter fitting may become much more problematic. For vector series, regression takes place on past values across locations, as well as the same locations. If the process is obviously not MVN the situation becomes considerably more challenging, although techniques do exist, as described e.g. in Scholzel and Friederichs [28].

One of the most common extensions to this model type is to introduce heteroskedasticity, i.e. non-constant variance, by allowing the random innovations to have a time-varying scaling factor. The scaling factor may be deterministic and designed to introduce seasonality to the variance, or may involve feedback, designed to introduce stochastic volatility clustering.

Another fundamental type of time-series model involve discreet Markov processes. For such processes the random variable representing the value of the series at time t, say X_t, may exist only in a limited number of states, usually defined by sub-divisions of the range of values it is observed to take. If the variable is supported over an infinite range, the top and bottom states may be open intervals. At the end of every time step X_t makes a transition - either into a different state or into the same state as it was in previously. The process is assumed to obey the Markov condition – i.e. it is memory-less, such that the state at time $t+1$ is conditional only on the state at time t. Sequences of transitions are known as Markov chains, their dynamics captured entirely by a transition matrix. The matrix elements p_{ij} represent the probability that the variable will make a transition from states i to j. In order to simulate a Markov chain, one must generate a random number for each time unit and use

this, together with cumulative transition probabilities, to determine the next state. In order to generate a continuous series an additional continuous, usually uniform, random variable is required to determine the position of X_t within the range defined by its state.

In the multivariate, i.e. vector, case the probability that a variable will be in state *j* at time $t+1$ is conditional on the states of each variable at time *t*, so a much larger matrix may be required – certainly the case for 20 dimensions. Higher order processes may also be defined, for which the memory-less condition is adjusted. For a 2nd order Markov process, for example, the state of variables at time *t*+1 is conditional on the state of variables at time t and $t-1$ only.

Discrete semi-Markov processes are a generalisation of the discrete Markov process. Instead of considering the transitions made at the end of each time step, for semi-Markov processes each state has an associated residence time – a random integer number of time units for series in discrete time. Assuming that the n^{th} transition is one from state i to state j after residing in state i for time τ_n, the probability that $\tau_n \leq t$ is a function of t, conditional on the state i only. It has the Markov property in the sense that the function of t does not depend on previous residence times or states.

Hierarchical processes may be defined which combine these two model types. An example is a process in which the observable quantity is defined by a regression process, but where the parameter values may suddenly change. These represent structural changes in the series and the period between changes may be considered regimes. The discrete state of the parameter values is a hidden random process, above the directly observable process in the hierarchy, and may have the Markov or Semi-Markov properties. Such a process seems intuitively to be a good choice in the case of wind speeds, where the regimes might represent meteorological conditions such as a low or high pressure system. In this case, the latter condition has the considerable advantage of allowing weather regimes with minimum residence times.

This research project takes the Bath Wind Model as a starting point, but will examine its validity and explore ways in which it may be improved. Possibilities include the development of a more complex model structure within the regressive framework, the development of a hierarchical model combining regressive and Markov-chain/ Semi-Markov aspects, or complete replacement with a Markov-chain/ Semi-Markov model, hierarchical or otherwise.

1.5. Modelling Justification - Power System Risk Assessment

The section addresses the essential question: why is there a need to produce synthetic wind data when long, consistent datasets are already available, e.g. from the Met Office? The main reason, as explained in detail below, is that power system risk assessments often require a large amount of data for very specific times, and historical datasets cannot provide this.

1.5.1. Context

The radical changes to the GB electricity system described above present many interesting and challenging questions relating to risk, which must be explored through probabilistic methods. Some of these questions are new, while others could previously be answered effectively using analytical methods. The biggest questions of interest are essentially: do the installed capacities of generators and interconnectors, combined with novel methods of system operation, ensure that the risk of electricity demands not being met is kept acceptably low? If inadequacy risk is increased are the associated financial risks, including those arising from a possible need to constrain some generators' output, acceptable?

A fundamental change to the nature of risks has occurred with the introduction of renewable generators. Conventional plants are often either 100% available or not at all, with well-known probabilities of being unavailable and no spatiotemporal correlations involved. None of this is true for renewable generators, however. For a traditional power system, being in a secure state could be equated with being able to continue operating and serve all demands following the sudden and unexpected loss of any single component, whether generator or network branch – a principle known as the $N-1$ criteria. This is too simplistic for a system with significant renewable generation capacity, as the generators' availability depend on weather conditions and vary continuously between zero and rated capacity. As stated in [29], probabilistic analysis is therefore the natural mathematical language for analysing systems with high renewable penetrations.

One of the most useful methods in probabilistic analysis involving power systems, and indeed the only one to be considered in this research, is sequential Monte Carlo simulation. Such simulations involve a large number of trials, corresponding to simulated points in time, for which samples are taken for the random variables of supply and demand (and possibly dynamic line ratings). Temporal order is maintained for such simulations and realistic temporal correlations are preserved – this is particularly important for future systems that include demand deferral and storage. The synthetic time series representing wind generator

availability generated by this research project are therefore ideal for use as one of many inputs in a full Monte Carlo simulation of a power system. The probabilities to be associated with forced loss of load events, or constrained generation, are then equated with their frequency of occurrence in the simulations. Such simulations may have to involve a very large number of trials, e.g. 100,000 in order to establish the precise probability associated with rare events.

This is a far superior approach to understanding the value of wind generation than commenting on individual events. Indeed, as stated by Gross and Heptonstall in [30], unless this type of large dataset analysis is undertaken the researcher is in danger of "moving back into the realm of the headline-grabbing fact selection that provides very limited insight into the power system implications of intermittent generation".

Power system risk assessments may be divided into two basic categories: those in which analysis takes place on planning and operational timescales. In the current context, a planning timescale may be conceived as one which is sufficiently far ahead of real time so that forecasts of demand and generation availability specific to the precise time under study are not possible. At any given time, the reality is that a subset of generators are deliberately offline or operating at limited output for economic reasons, and calculations of adequacy on operational timescales make use of this information. When working on planning timescales, however, it must be assumed that every MW of generation capacity is available if it can be, i.e. not prevented due to essential maintenance, fault or renewable resource availability.

It is in the planning timeframe that the adequacies of generation and transmission capacities are assessed. An important example is the assessment of the adequacy of a present generating fleet, given the calculated underlying demand level for a coming peak season. Alternatively one may examine the validity of future energy system scenarios, such as those developed for the LENS project. As previously stated, these are usually developed using purely economic models and require validation that accounts for the spatiotemporal characteristics of supply minus demand.

Many risk assessment calculations simplify the situation by concerning themselves only with the system-wide balance of available generation and demand, ignoring all other factors such as network capacity constraints, the optimal use of hydro generation etc. As stated in [31], these simplified calculations are described as Hierarchical Level 1 by the relevant Institute of Electrical and Electronics Engineers (IEEE) Task Force [32]. The multivariate nature of the series generated in this project mean that a researcher may

certainly move on to more advanced studies where the balance of supply and demand within regions, and the extent of required power flows between them are calculated. The datasets could potentially be used in calculations involving any number of the additional factors, if suitable models or data to represent such factors were available, and the meteorological implications of wind speed patterns were sufficiently well understood. It must also be noted that simplified methods are not necessarily inferior to more complex ones - simpler calculations may be more transparent, for example, providing more insight into what drives the results, as discussed in [33].

One area of application where network effects are of central importance is the evolution of GB's transmission planning standard. As stated in [29], transmission and distribution networks have traditionally been designed according to deterministic network planning standards that make use of relatively simple heuristic rules, designed to give a reasonable solution to a complex underlying probabilistic planning problem. The GB 'deterministic' planning standard is somewhat unusual in that it provides a formula for the total required transfer capability across any boundary between groups of zones. As GB moves towards a future network with a much higher penetration of renewable generation, such rules must be revised using probabilistic methods. Aiding in this specific process was the main motivation for developing the multivariate Bath Wind Model.

1.5.2. Probabilistic Calculations and Insight

Electricity demand during a given period will never be exactly the amount predicted, and there is always a probability that generators cannot supply the amount promised. This means that at any point in time there is always a small probability that supply will not meet demand – this is the loss of load probability (LOLP) for that instant, an hour in the current context. Dent and Zachary in [35] choose a random variable M to represent the excess of supply at some instant, and note that its expectation value is always significantly greater than zero - although much more so on e.g. summer nights than winter evenings. If the cumulative probability distribution function (CDF) for M at the time of interest is $F_M(x)$, then the LOLP is $F_M(0)$, a very small number for a single hour.

Whereas a LOLP value is associated with a 'snapshot' in time, adequacy studies are concerned with extended periods, typically the entire year, where the relevant probability distributions vary from day to day and hour to hour within that period. The measure of generation adequacy risk for the period is therefore the sum of the LOLPs associated with the

individual time periods (hours) within the longer period – known as the loss of load expectation (LOLE), a (non-integer) number of hours in the present context. If this summation comes to e.g. 0.15 for a year, we may express this as an expectation that there will be 3 hours in every 20 years during which not all demand can be met due do an insufficient amount of built generation capacity, or insufficient transmission capacity to carry power from generators to the locations of demand. There may in reality be a greater number of unserved demand hours, due to e.g. faults at local substations or distribution lines falling after a heavy snow.

In some contexts, e.g. [25], the term LOLP is used in a different, rather inaccurate way – as a percentage that is the number of years per century in which supply is insufficient at any point. It may be the case that the intended meaning is that demand is unmet during one period only per winter, on average, but it is not a very clear proxy for a probability value. Regardless of the manner of definition, probability metrics say nothing about the distribution of durations for forced load-shedding events.

An alternative adequacy metric is sometimes employed [30]: the loss of energy expectation (LOEE), i.e. the expected number of unserved MW-hours per year due to lack of generation or transmission capacity. As it is more thorough, calculations involving the LOEE may be considerably more technically complex. However, the result is more straightforward to interpret.

A very common type of risk based calculation carried out for wind generators during recent years is that of capacity credits (e.g. [30]). Many definitions of capacity credit exist, with two more common, but they all relate to the marginal properties of a generator, or set of generators, when added to a specific background of generation, demand and possibly network [35]. One of the most popular definitions is Effective Load Carrying Capacity (ELCC): the additional deterministic (i.e. constant) demand which the additional capacity can support without increasing risk, usually as captured by the LOLE. The other popular definition is Equivalent Firm Capacity (EFC): the deterministic capacity whose addition would give the same decrease in risk, usually as captured by the LOLE, as the additional capacity in question. Zachary and Dent show in [35] that the two credits are almost equivalent for relatively small capacity additions. Arguments can be made for the conceptual superiority of either definition, but this seems to depend on the context.

Following [35], the availability of additional capacity of interest (in the present case wind), is represented by the random variable Y.The analysis assumes that the mean and variance of Y, μ_Y and σ_Y^2, are both small compared to those of the entire set of generators. The

assumption is also made that Y and M are independent, and the probability density function (PDF) for M is written as $f_M(x)$. Dent and Zachary show that to a good approximation the capacity credit (using either ELCC or EFC) is given by

$$v_Y = \mu_Y - (f'_M(0)/2f_M(0))\sigma_Y^2. \quad (1.2)$$

The equation demonstrates clearly that capacity credits are only well defined in relation to the probabilistic background created by the pre-existing supplies and demands. Also, it demonstrates that only conditions at the point of zero margin are of interest, as intuition might suggest. In the much more reasonable case that Y and M may be dependent, the mean and variance must be replaced with their values conditional on $M = 0$.

If M is decomposed into two random variables according to $M = X - D$, where X is available conventional generation and D is demand, then LOLE calculations require knowledge of the joint distribution of X, D and Y – generally a very 'messy' situation, mathematically. Fortunately, it is often a reasonable approximation to treat X as independent of the joint distribution of D and Y. In this case, as shown by Dent and Zachary, the situation becomes simpler with e.g. the conditional mean of Y given by

$$\mu_{Y\,|M=0} = \int_{\mathbb{R}} \mu_{Y\,|D=x} f_X(x) f_D(x) dx \Big/ \int_{\mathbb{R}} f_X(x) f_D(x) dx. \quad (1.3)$$

For the current GB system, X has a much narrower distribution than D, i.e. the left hand tail of X decays more rapidly than the right hand tail of D, which means that the calculated LOLE will be dominated by times of extreme demand, taken here to be winter 5pm – 7pm. In other words, it is extremely unlikely for the current system that M will be below zero unless demand is extremely high. Equivalently, the product from the equation above, $f_X(x)f_D(x)$, is only significant when X is near peak demand – so capacity credit can be conceived as the additional capacity's ability to provide support during times of peak demand. This definition would be too narrow if the distribution of X was relatively broader, as is the case for smaller power systems, and also if one was interested in the capacity credit of additional wind capacity in a future system which already had a high penetration of renewables, and possibly measures to limit demand peaks.

When calculating the LOLE, it is usually sufficient to concentrate on particularly "risky" hours or days, over which the relevant random variables may be approximated as being identically distributed. The 'snapshot' LOLP for this period then becomes a substitute for the full LOLE. The fact that calculations involving LOLE require knowledge of the relevant joint probability distributions illustrates the need for appropriate data – i.e. many repeated

concurrent values. These should ideally be of the triple (X_t, D_t, Y_t), otherwise the pair (X_t, Y_t), observed over a sufficient number of periods for which they may be considered identically distributed.

1.5.3. Hindcasting vs. Sequential Monte Carlo Simulations

The most popular method for the practical LOLE calculation is known as hindcasting (e.g. [31] and [34]). In this technique conventional generators are treated as independent, with time-invariant Bernoulli distributions. This implies that X has a binomial distribution formed by convolution of the Bernoulli distributions. Introducing a time subscript, the random variables D_t and Y_t are however replaced by a historic time series of their concurrent values. This makes calculations significantly simpler – for capacity credit calculations using the EFC definition, for example, the deterministic demand v_Y^{EFC} may be calculated by solving for it in the relatively simple equation

$$\sum_t F_X(d_t - y_t) = \sum_t F_X(d_t - v_Y^{EFC}). \quad (1.4)$$

The summation here is over all hours in the historical time series, which should be a large number, allowing an appeal to be made to the Central Limit Theorem to justify the validity of this approach.

Detailed meteorological records generally extend back around 100 years, corresponding to about 877,000 sample hours. However full and consistent records for specific locations tend to be no longer than a few decades – and thus are not necessarily entirely representative of the resource. It has been established above that LOLE's are dominated by hours of peak demand, well known to occur between 5pm and 7pm in the months of December to February (e.g. [26], [7]). Considering only such hours, a 20 year sample is reduced to 3,620 hours, which rather small for the evaluation of joint probabilities.

Considering actual occurrences of truly extreme values, the situation becomes significantly worse. When discussing peak demand, a very useful concept is the average cold spell (ASC) peak, the demand level that has a 50% chance of being exceeded during a winter period due to weather conditions alone. In another paper [36], Dent and Zachary examined a coincident time series for transmission-metered wind load factor and demand in GB for the four winters between 2006 and 2010, finding that the number of hours above 95% ACS peak was only 115, far too small to permit a robust statistical analysis. In [31], Zachary, Dent and Brayshaw examined data from the nine year period since 2001 for which GB-wide demand data is available. They discovered that there were only 18 days during which 99% of ACS peak

was exceeded, and that two thirds of those days were in just two periods, in the Januarys of 2009 and 2010 respectively. This is clearly insufficient to give any statistical picture whatsoever of the wind resource at the relevant times.

England and Wales demand for a longer period of 20 years was also available to Zachary, Dent and Brayshaw, which may be taken as a good proxy for GB demand. They examined Very Cold Spell (VCS) peaks, now preferred by NG over ACS, and about 2% higher. Again all periods of such demand came from a small number of distinct periods: 5 days, all in a single weather system in January 1987. They note that these particular days would dominate any adequacy calculation, and must be regarded as a type of event to which one cannot yet presently assign an accurate probability. Studying the types of weather system which have historically driven such extreme demand events is presented as a means of overcoming, to an extent, these limited data issues.

Further evidence comes from the research project at Edinburgh University mentioned in section 1.3.1 [19]. Having created a 10 year dataset of onshore and offshore wind speeds, Hawkins and colleagues went on to use it in hindcast calculations of the capacity credits of future GB wind fleets [19]. Examining demand data for the 10 year period, the authors found that the number of hours with demand in the range 90 – 94% ACS peak was 1,656 – a reasonable number, but for the range 95 – 99% ACS peak there were only 447 hours, and only 40 hours for the range 100 – 103%.

Hawkins et al. [19] developed a specific hindcasting methodology to calculate capacity credit, which is a fully valid contribution to current approximations of capacity credit in GB. This is believed to be the first time that capacity for a combined on- and offshore wind resource has been calculated using such an approach. They acknowledge however that 10 years is not enough to represent a full climatology, despite being long enough to sample a wide range of synoptic conditions, which is an improvement on many studies. They note that the output of wind at times of peak demand varies considerably between years, and that this “highlights the difficulty, perhaps even the validity, of attempting to represent the contribution wind makes towards reliability as a single figure”.

In other words, a major drawback of the hindcasting approach is that it gives no idea of uncertainty in values derived through it, since it is not an explicitly probabilistic approach. As discussed by Dent and Zachary in [36], boot-strapping techniques may be used to provide such estimates through resampling of the historical dataset, but this relies upon assumptions such

as that successive values for wind availability and demands are uncorrelated – clearly not a good approximation.

The alternative approach, as discussed in Section 1.5.1, is to use Monte Carlo simulations with synthetic wind datasets, a probability distribution or time series model for demand and some stochastic representation of other variables**.** If only winter evening values are extracted from the original wind series, a sample dataset of just under 100,000 such hours could consist of 55 different decade-long realisations of the stochastic process, with each one using the same historical data for starting values. Such a dataset could yield 55 capacity credit values from decade-long simulations, providing a rough distribution for the credit value. This clearly represents a thorough exploration of 'what could happen' in terms of the resource.

A significant downside to Monte Carlo simulation is that, unlike hindcasting, one does not automatically reproduce exactly the joint distribution of wind availability and demand. In principle one could extend the time series model to include wind availability and demand, but capturing such complicated relationships would be extremely challenging, and the simplifying assumptions might defeat the object of capturing extremes. Whilst the use of a synthetic time series permits assessment of distributions for the frequency and duration of supply shortages, without representation of the joint distribution with demand, using a synthetic time series for wind will give results equivalent to sampling from the corresponding marginal distribution. This is not true for simulations of future scenarios in which there is significant storage or load deferability, however.

The challenge for this project is therefore to develop a model accurate enough so that a distribution of values obtained via simulations seems preferable to the equivalent single value obtained through a hindcast calculation. Unfortunately there is no obvious way of proving which approach is the best.

1.5.4. An All-Year vs. Winter-Only Model

The importance of peak demand times for many applications might motivate a decision to fit models specifically to winter data, which might increase their accuracy for this season. However this would limit the scope of the model's applications. For example, for a future scenario with a very high penetration of wind power, rigorous evaluation of the scenario should include the extent to which wind would be curtailed during the summer due to proposed set of network capacities. Additionally, a possible future paradigm shift in the use of

micro-generation and low carbon technologies could significantly broaden the times of interest to risk studies [7].

Another reason to develop a model for the entire year is that they may be useful in operational timescale studies. An example area of interest on this timescale is the system operator (SO)'s reserve setting standard. This is a risk-based standard for determining the 4-hour-ahead Short Term Operating Reserve Requirement. This requirement relates to the amount of reserve needed to handle unpredicted short term variations – traditionally either due to demand prediction errors or generation failures, but now also due to renewable generation prediction errors. As discussed in a highly authoritative report by the UK Energy Research Centre on the effects of renewable generator intermittency [37], this may be worked out through analytical techniques using statistical principles, or through Monte Carlo simulation models. It is noted that analytic techniques provide approximate results whereas simulations are needed to deal with complex, realistic situations.

1.6. Other Applications of the Algorithms

1.6.1. Forecasting

The simulation algorithms will also be capable of generating wind speed forecasts in real time. Due to the sensitivity of meteorological models to initial conditions throughout the atmosphere, which cannot be known with complete accuracy, time series models can outperform meteorological forecasts for horizons of up to 3 or 4 hours [38]. However, the model developed here will be optimised for long simulations – those designed for forecasting would incorporate real time information such as wind direction and pressure.

The regression models described above have two components – one deterministic and the other stochastic. In addition to being a temporal filter, the deterministic component is also a one-step ahead forecast for the series, with the stochastic innovation term representing the forecast error. This means that for one hour ahead forecasting, regression models will not only yield a point forecast very easily, but the probability distribution for the forecast is also easily given. For longer forecast horizons further, both the point forecasts and approximate distributions may be obtained by generating a large number of series 'runs' up to the time in question, using recently recorded wind speeds as starting values each time. The relative simplicity of time series models allows the number of runs to be high without creating an excessive computational burden. In the case of models such as Markov or Semi-Markov chains, the forecast is a weighted average of all possible transitions for 1 hour ahead, and must again be calculated based on numerous runs for longer periods.

While it is worth noting the potential use of the model and algorithm for forecasting in this way, an investigation of forecasting ability is beyond the scope of this project.

1.6.2. Making Data Available to All

Historical wind datasets collected specifically for wind energy purposes are scarce, partially because wind energy has emerged only fairly recently as a mature and large scale industry. Data from realistic locations has only been collected by wind farm developers, usually covering a period of a few years, and the data is considered commercially sensitive. Monthly averaged load factors for individual wind-farms are publicly available – published indirectly by Ofgem and processed by the Renewable Energy Foundation [39]. These are useful, but limited by the very heavy temporal aggregation. It is also possible to get high temporal resolution data for wind generation outputs from market settling company Elexon [40], but the problem in this

case is total spatial aggregation. Fortunately, hourly resolution and location specific data has been collected by the Met Office over several decades - although locations are far from ideal. Since these are available to registered academic users only, there is value in making synthetic wind data trained on the Met Office data publicly available, along with synthetic power outputs for an example scenario. The datasets may even be integrated into an open source power system or renewable energy project such as OSeMOSYS [41]. The Author is aware, through personal communication, that National Grid aspires to produce and make available synthetic wind datasets trained on commercially sensitive historical data from actual wind-farm locations. By establishing the best type(s) of model for synthesising wind time-series, this research will surely contribute to such endeavours.

1.7. Climate Change

A major assumption in this research is that the wind resource during the last few decades, and to a lesser extent the last century, is representative of the future wind resource. As will be discussed in detail in Chapter 2, the wind climate is constantly changing. This means that the future resource will not be identical to the past, and this should be reflected appropriately in the synthetic series. However the assumptions made here are that the ways in which it will change are unpredictable, and that the extent of differences between e.g. decades will remain consistent with changes observed in historical datasets.

This appears to be potentially at odds with the premise that the Earth is currently at the beginning of a period of relatively rapid climate change resulting from the release of greenhouse gases by human activity. This is quite a pertinent issue since climate change is one of the main motivations for delivering the radical decarbonisation of the energy system. However, a clear message within authoritative texts such as [42], and reports by the Intergovernmental Panel on Climate Change [43], that there is no scientific consensus on any predictable changes to the GB wind resource. Indeed, confidence in future changes in windiness seems to remain relatively low, although it appears more likely than not that there will be an increase in average and extreme wind speeds in northern Europe. Some studies point in the opposite direction, however, and areas of strong orographic forcing (e.g. steep valleys and ridges) could see major differences or even changes in the opposite direction to the general large scale behaviour.

Reports do however confirm the well-known confidence among scientists that GB will experience warming over the coming decades, greatest in winter, although probably countered in part by a likely reduction in the warming effects of the Gulf Stream. It is believed that by the mid-21st Century inter-annual summer temperature variability is likely to increase, and summer 'heat waves' are very likely to increase in frequency, intensity and duration (although less so than continental Europe), resulting in a much increased cooling load. In contrast with summer, models project reduced temperature variability in most of Europe in winter, both on inter-annual and daily time scales – corresponding to more predictable winter peak loads.

A preliminary investigation into the possible effects of climate change on the GB wind resource was conducted by Harrison, Cradden and Chick, as reported in [44]. Their research is based on wind speed projections taken from regional scenarios published by the UK Climate Impacts Programme in 2002. They provide mean-monthly changes for a range of climate

variables on a 50 km × 50 km model grid for three periods or 'time slices' representing conditions for 2011–2040, 2041–2070 and 2071–2100 respectively. Projections were made for four scenarios of future greenhouse gases emissions - the 'high' scenario for example representing a situation where emissions are more than triple current levels by 2050. The authors find that for this extreme scenario, GB could potentially experience 5–10% increases in winter wind energy production, despite a slight reduction in production in the north of Scotland. Summer changes appear to indicate lower summer turbine capacity factors, falling by up to 10% - although some areas would experience more severe reductions.

It is worth noting that the traditionally popular term 'global warming' might be better replaced by the term 'global weirding'. As popular US journalist Thomas L Friedman states, "I prefer the term 'global weirding', because that is what actually happens as global temperatures rise and the climate changes. The weather gets weird" [45]. Global weirding is a phenomenon that is already very much apparent, and may well be responsible for a very poor resource in the winter of 2010, and to a lesser extent winter 2009, behaviour directly opposed to the predictions described above. It seems therefore that none of the work described here suggests that the assumption that records for the last few decades can be used to explore the value of wind in the coming decades is unreasonable.

1.8. Research Objectives and Structure of the Report

This first chapter has already thoroughly explored the motivation for producing synthetic datasets of the wind speed field, time-stamped and covering the entire year. It has placed this research project within both an academic context and a socio-political one. It has also provided a concise yet complete sweep of the types of time series models that must be explored and considered before establishing the optimal model structure. This section sets out the remaining objectives, which evolved during the course of the research, matching them with the corresponding section of the report.

Chapter 2:

Conduct a literature review on the nature of the wind resource, with a focus on GB.

2.1: Conduct a literature review on basic characteristics of the wind resource, particularly in GB – an area of study known as wind resource meteorology.

2.1.1: Provide a description of the sources of wind, particularly in GB, with reference to synoptic scale variability and diurnal seasonality.

2.1.2: Review the suitability of the hourly resolution.

2.1.3: Present and discuss wind speed distributions.

2.1.4: Present and discuss wind shear – theory and recent analysis.

2.2: Provide a literature review of the spatiotemporal properties of the wind resource – both in terms of wind speeds and wind farm outputs.

2.2.1: Present and discuss the work of Dr Graham Sinden on the GB wind resource.

2.2.2: Present and discuss relevant technical reports by the United States Government's National Renewable Energy Laboratory (NREL).

2.2.3: Discuss the relevance of atmospheric energy conservation principles.

2.3: Conduct a literature review of the meteorological practice of synoptic classification.

2.3.1: Provide an introduction to concepts and relevant literature.

2.3.2: Provide a detailed description of the GWL synoptic classification system and its relationship to peak demand times. Also the HB-GWL circulation type catalogue, re-analysis datasets, and a reduced set of GWL types.

2.3.3: Provide a literature review on the objective classification of surface wind measurements.

2.4: Provide a meteorological literature review on the nature of climatological variability.

2.4.1: Provide a description of inter-annual and inter-decadal variability in mean wind speed and mean wind power load factor.

2.4.2: Provide a meteorological literature review of long-term spatiotemporal patterns in atmospheric circulation in the UK and Europe.

2.5: Discuss alternative mathematical approaches for representing climatological non-stationarity.

2.5.1: Discussion of the concept of long memory in time series, and its presence in meteorological phenomena.

2.5.2: Discussion of heteroskedasticity and its presence in wind speed time series.

Chapter 3:

Present the fundamental mathematics of time series analysis and modelling, providing a reference for future chapters. This includes the fundamental tools of analysis, the various types of model structure commonly used and model fitting techniques.

3.1: Present the fundamental principles of ARMA Models.

3.1.1: Present the basic concepts, notation and terminology associated with time series analysis and ARMA models – particularly correlograms and partial-correlograms.

3.1.2: Discuss spectral the representation of time series, from periodograms to the transfer functions of ARMA models.

3.1.3: Present the theory of SARMA seasonal models, including their ACF and spectral densities.

3.1.4: Present the concepts associated with differencing time series, including when it's necessary. Introduce ARIMA and SARIMA models in their most general form.

3.1.5: Introduce transformation methods, including Box-Cox and those involving inverse cumulative distribution functions.

3.2: Present ARMA model extensions that allow representation of long memory and heteroskedasticity.

3.2.1: Present extensions to ARMA models that allow for long memory – both 'regular' and seasonal, i.e. ARFIMA and GARMA. This must include extensions to the multivariate case, and how fractional differencing may be achieved in practice.

3.2.2: Present univariate models for conditional heteroskedasticity: ARCH, GARCH and further generalisations – including APARCH and ARCH-in-mean. This must include possible extensions to the multivariate case.

3.3: Presentation of the theory and practice of order selection and parameter fitting for ARMA type models.

3.3.1: Provide a conceptual introduction to the fitting of ARMA models.

3.3.2: Present the Yule-Walker equations and the Levinson-Durbin algorithm. Also, Whittle's multivariate extension.

3.3.3: Present and discuss the role of information criteria (the AIC and BIC) in model order selection, including as part of the Hannan-Rissanen procedure.

3.3.4: Present the theory of maximum likelihood estimation, including the multivariate case.

3.3.5: Present a series of tests that may be applied to model residuals to check if model structure is adequate.

3.4: Presentation and discussion of the theory of wavelet transforms, and their potential use in modelling the wind resource, particularly if a hierarchical model structure is chosen.

3.5: Discuss non-Gaussian multivariate processes, and the use of copula functions as a means of going beyond the multivariate normal assumption.

3.6: Presentation of Markov Chain Monte Carlo models, with discussion of their potential benefits. This includes extensions to the basic model type, including higher orders and multivariate models.

Chapter 4:

Prepare a second literature review, reporting on efforts to date to create and apply time series models to wind speeds. This includes previous work at the University of Bath that acts as a starting point for the model developed here. Drawing upon chapters 2 and 3, the most advanced models should be discussed and critically evaluated.

4.1: Presentation and critical evaluation of previous wind speed modelling work conducted at the University of Bath.

4.1.1: Present and critically evaluate the Bath wind model in greater detail than Chapter 1, including conversion of the wind speeds initially generated to zonal power outputs.

4.1.2: Critically examine the way in which the Bath wind model's performance and suitability was previously assessed.

4.1.3: Critically review other wind modelling work conducted at the University of Bath.

4.2: Review ARMA/VARMA models for wind speed developed by other authors.

4.2.1: Review work that constructs univariate ARMA models for wind speed.

4.2.2: Review work that constructs VARMA or similar multivariate models for wind speed.

4.3: Review work modelling wind speed as a Markov Chain, or a semi-Markov process.

4.4: Review work on hierarchical models of wind speed.

4.4.1: Review work on regime-switching models of wind speed.

4.4.2: Review the use of wavelet transforms in wind speed field modelling.

4.5: Review the literature on producing wind power outputs from wind speed time series.

4.6: Review the literature on the direct modelling of aggregated wind power outputs.

4.7: Review work that models wind speed time series as possessing long memory and conditional heteroskedasticity – ideally models that incorporate both.

4.7.1: Review work where wind speeds are modelled as an ARFIMA or a GARMA process.

4.7.2: Review work where wind speeds are modelled as having conditional heteroskedasticity.

Chapter 5:

Explore available wind observation data, choose the best set of meteorological station for each zone, assess of their suitability as locations and improve data quality.

5.1: Acquire example wind speed observation data, and convert to a format that's easy to manipulate in *Matlab* (the software choice for almost all the calculations involved in this research). Briefly assess its quality, including visual inspection of time series segments.

5.2: Clean-Up the Datasets – identify what should be done, and developing algorithms to do it.

5.2.1: Develop an effective clean-up algorithm to ensure correct chronological ordering, to identify gaps and to choose between multiple entries for the same hour.

5.2.2: Present and characterise wind speed distributions

5.2.3: Explore options for the removal of erroneous readings of 0 knots.

5.3: Choose the best combination of Met Office stations and assess the resulting quality of data.

5.3.1: Explore many MIDAS observation stations to find the best choice of 20. This must br on the basis of multiple criteria – some relating to the individual stations and others to the entire set.

5.3.2: Asses and report on the quality of data for the final selection, including the distribution of gap lengths.

5.3.3: Use *Google Maps* and *Google Earth* to characterise the precise locations of the Met Office station masts.

5.4: Fill-in all gaps in the data (as some Matlab functions require this). As part of this, transform the series so that they roughly form a multivariate normal, with standard normal marginal distributions.

5.4.1: Power transform the series – after exploration and consideration of the optimal way of doing so. Explore the effectiveness of the transformation.

5.4.2: Develop an algorithm to remove deterministic diurnal seasonality in mean and variance.

5.4.3: Establish a method for filling-in the gaps, based on the exploration of several interpolation and forecasting approaches.

Chapter 6:

Conduct a detailed statistical analysis of the historical series, hopefully confirming and enhancing insights from chapter 2. Explore the usefulness of transformation of the series into principal components and explore how the compoents relate to the GWL atmospheric circulation classification scheme. Explore the potential use of relationships found.

6.1: Conduct a statistical analysis of the wind speed series, with diverse aspects of wind resource dynamics examined.

6.1.1: Present and comment upon summary statistics for each series, along with their cross-correlations. Particular attention must be paid to skewness coefficients – before and after power transformation, as a reflection of how similar the series' distributions are to the (standard) normal.

6.1.2: Prepare and discuss plots of time series segments, with wind speeds averaged over vastly different time scales.

6.1.3: Prepare and discuss figures exploring differences between months, in terms of summary statistics. Scatter plots may be used to simultaneously present the relationship between 2 statistics.

6.1.4: Prepare and analyse a selection of correlograms and partial-correlograms, with an emphasis on exploring a longer range of lags than typical presentations found in the literature.

6.1.5: Prepare and analyse a selection of periodograms, some with logarithmic axes.

6.2: Perform principal components analysis on the 20 series; present and discuss the results.

6.3: Match a daily catalogue of reduced GWL circulation type to wind speeds across the 20 year period, and analyse their relationship.

6.3.1: Examine the 20 year period in terms of changes in the relative frequency of occurrence for the reduced GWL types.

6.3.2: Estimate probability distributions (i.e. relative frequency of occurrence in the historical period) for the principle components for each of the GWL types. Discuss any differences found.

6.4: Investigate thoroughly the clustering of wind speed fields, and whether the clusters correlate with certain weather types – or even correspond to a significantly higher/lower probability that a certain weather type is occurring.

6.5: Explore whether the relationships established might suggest a potential method of connecting electricity demand and the wind speed field is proposed.

Chapter 7:

Develop and apply an iterative process for choosing the model structure, establishish optimal parameter values, testing the suitability of those choices and considering possible improvements to the model structure.

7.1: Establish and apply the best method for fitting the initial choice of model structure for the conditional expectation values: an annually seasonal VGARMA model. Use maximum likelihood estimation (MLE) if possible.

7.1.1: Develop and apply a methodology for fitting an annually seasonal VGARMA model to the 20 series, initially making use of techniques such as the Hannan-Rissanen procedure, rather than (MLE).

7.1.2: Provide a detailed literature review of attempts to use MLE, and simpler maximum quasi-likelihood estimation, for model structures similar to VGARMA.

7.1.3: Attempt to use quasi-MLE to establish the parameters of the VGARMA model, using those previously established as starting values. Describe the relative success of the method.

7.2: Discuss why the choice of an annually seasonal VGARMA model structure was replaced by a 2-Factor-VGARMA model with deterministic annual seasonality. Develop and apply a methodology for fitting the parameters of the mew model structure without use of quasi-MLE.

7.2.1: Present reasons why the annual seasonality is probably best modelled as deterministic. Develop and apply a methodology for identifying and removing this annual seasonality, making use of smoothing to an optimal extent.

7.2.2: Prepare and analyse periodograms, smoothed to varying extents, with the deterministic annual seasonality removed. Describe how they provide evidence of the suitability of a 2-factor-VGARMA model structure.

7.2.3: Develop and apply an optimal methodology for establishing the parameter values of a 2-factor-VGARMA model fitted to the data. The possibility of using quasi-MLE must be investigated as part of this process.

7.3: Conduct detailed analysis of the residuals of the 2-factor-VGARMA model, with a focus on spatiotemporal associations. Discuss the implications of the analysis in terms of a proposed model structure for conditional variance. Develop and apply an optimal parameter fitting process for the conditional variance model.

7.3.1: Conduct detailed analysis of the 2-factor-VGARMA model residuals, particularly their spatiotemporal structure, with the main tool being autocorrelation functions.

7.3.2: Develop and apply an algorithm to remove observed itra-annual and inter-annual seasonality in variance, leaving it as constant as possible.

7.3.3: Discuss the suitability of univariate APARCH models to capture the remaining volatility clustering behaviour. Establish whether a sub-set of such models is appropriate, to simplify parameter estimation. Develop and apply a methodology for estimating the model parameters. Assess their success in removing spatiotemporal associations in variance.

Chapter 8:

Develop and apply a complete simulation methodology for the fitted model, starting with suitable i.i.d. noise and adding layer upon layer of structure until finally wind speeds are produced. The development of the methodology must involve assessment of the simulation results at each stage of the process, leading to changes to the model structure if appropriate. Compare diverse aspects of the simulated series to those generated from a simpler model, using the historical series as a reference. Develop a realistic scenario for the distribution of wind capacity in about a decade, then convert simulated wind speeds to aggregated wind powers for this scenario.

8.1: Develop and apply a set of algorithms for generating the Synthetic Series. Analyse the model's performance at each stage and use this as the basis for making adjustments to the model structure, if appropriate.

8.1.1: Develop and apply an algorithm capable of generating temporally independent random deviates that have the same spatially-joint distribution as the unconditional residuals from the historic series.

8.1.2: Develop and apply an algorithm for generating conditionally heteroskedastic noise series with temporal structures given by the fitted APARCH model. This may be more involved than simply applying the APARCH model structure to the unconditional noise series. Also, develop a simple algorithm for re-establishing the inter- and intra-annual patterns in variance.

8.1.3: Develop and apply algorithm for constructing wind speeds from the final synthetic noise. The algorithm must filter the noise with the VARMA model then reverse the double fractional differencing.

8.1.4: Develop and apply an algorithm for making final adjustments to the wind speed series, which may involve appropriate censoring. Principles of appropriate censorship must be established. In addition to matching the synthetic and historical series in terms of mean and variance, kurtosis must also be realistic, with an algorithm developed to ensure that this is the case. Develop and apply a simple algorithm to apply the last reconstruction stages, including reverse power transformations.

8.2: Conduct an initial analysis of the synthetic series.

8.2.1: Calculate and discuss synthetic wind speed distributions at single locations, comparing with historical distributions.

8.2.2: Calculate and discuss plausible wind power distributions from the wind speeds – at single locations and for each zone aggregated (all zonal weightings equal). Apply more than one up-scaling factor for generality. Calculate and discuss the distribution of changes in the aggregate power, for 1 hour and 4 hour time differences, historical and synthetic series.

8.2.3: Develop and apply a methodology that allows examination of the distribution of multivariate states characterising in the wind speed field, allowing comparison of historical and synthetic joint distributions beyond simply the aggregated marginals. Discuss the results.

8.3: Fit a multivariate transition matrix model to the doubly-differenced series at a subset of zones. The model should 'nudge' the VGARMA-APARCH model in such a way as to improve spatial joint distributions. Analyse whether this additional modelling step has enhanced simulation quality.

8.4: Conduct thorough analysis of the dynamics of the synthetic and historical series, to further the extent of their similarity. Develop and fit a simpler (yet good quality) time series model to add to the analysis – to establish whether the complexity of the 2-factor-VGARMA-APARCH is justified.

8.4.1: Calculate and discuss autocorrelation functions, ranges 0 – 500 hours and 1001 – 10,000 hours, for the historical series and those generated from the two models.

8.4.2: Calculate and discuss several periodograms, for the historical series and those generated from the two models, focusing on different frequency ranges.

8.4.3: Conduct and discuss analysis on inter-annual variability and long-term means, for the historical series and those generated from the two models. The analysis must consider: the variance of means, for January, July and annual means; long term means (for the same 3 periods); long term variances (for the 3 periods); and the variance of monthly variances (for the 3 periods). Compare the correlation between monthly means and variances. Examine the extent to which the two synthetic series represent the wide range of monthly skewness values found in historical series.

8.5: Develop a realistic wind capacity scenario for the fairly near future. Procude aggregated power outputs for this scenario from the synthetic series (generated using the more sophisticated model). Assess how realistic is the aggregated power output series.

8.5.1: Develop and apply a methodology to represent diurnal seasonality at hub height accurately. Develop and apply a simple approach for speeding-up wind speeds from the

recording station locations to realistic wind farm locations, also speeding-up from 10m to hub height.

8.5.2: Develop a realistic scenario for an unspecified time in the fairly near future, specifying the amount of capacity at each type of location within each zone. Develop a policy on what type of capacity may be excluded when constructing the scenario.

8.5.3: Develop and apply an algorithm for smoothing the single-location wind power outputs adequately, so that the aggregated output has realistic dynamics. Develop criteria for testing whether this is the case. Provide example segments of the aggregated series.

1.9. The Unique Contributions of this Research

This research project is the first to investigate thoroughly the best choice of vector time series model structure for the long-run simulation of the wind speed field across a large-area power system. It does so for GB, where the wind resource is excellent and investment in wind energy may be very significant over the next decades.

Previous research on characterising diverse aspects of the wind resource (in GB and internationally) is presented, discussed and built upon - including means of characterising very low frequency variability and means of classifying the overall meteorological 'situation' over GB.

New developments at the frontier of time series modelling are explored and exploited in order to arrive at a model structure with many novel aspects. This research project is the first to apply multivariate GARMA to wind, and is apparently the first attempt, in any context, to combine GARMA and conditional heteroskedasticity in a multivariate model. All of the previous work described focuses on the recreation of some aspect of the wind resource's dynamics in detail, such as getting the multivariate relationship correct, the long memory, diurnal seasonality or the smoothing effects of multiple turbines. None consider all such aspects with much sophistication, since complexity renders this unfeasible. However this research achieves compromises that give at least reasonably good consideration to all relevant aspects. It remains possible, however that a simpler model structure might be equally successful at achieve such a balance.

This research project makes a valuable contribution to a vast and currently immature field: understanding the best ways to conduct probabilistic analyses of future power systems, where the sequence of events matter, including adequacy assessment and capacity credit assessment.

Chapter 2

The Wind Resource

2.1. The Foundations of Wind Resource Meteorology

This chapter begins with a review of the basic facts about wind resources, specifically in GB where appropriate, as presented in wind energy texts such as [50], [51], [52] and [53]. This field is referred to in [52] as wind resource meteorology – a combination of meteorology, climatology and geography.

2.1.1. Sources and Nature of Winds in GB

Wind is the movement of air resulting from parts of the earth's surface being heated by the sun to a greater extent than other parts. The atmosphere contains motions with scales varying from about 1 mm to thousands of kilometres. On regional scales winds may consist of sea breezes caused by differences in surface temperature between sea and land arising during the course of a day. On very local scales are thermals - columns of rising air created by air close to the surface being heated to a much greater extent than the air above it. However in GB the wind resource is dominated by what is known as synoptic activity: the passing of dynamic low and high pressure systems hundreds of km in extent.

The atmosphere at the surface of the earth consists of a number of very large air masses - volumes of air covering many hundreds or thousands of square miles, defined by their roughly homogenous temperature and water vapour content. The masses adopt the characteristics of the surface below them, so continental air masses are dry while maritime air masses are moist. Where two air masses meet their often very different characteristics, mainly temperature and density, prevent them from easily mixing and instead they form a 'front'. The weather at a location such as GB is defined to a large extent by the passage of such fronts.

The excellent GB wind resource is a result of phenomena on the western side of the North Atlantic, as described by Oswald, Raine and Hezlin [54]. In this area, a warm air mass from the tropics moves north until it meets cold air moving south and east off the Canadian land mass. These air masses are several times larger than a European country, and collide at

approximately 40^{o} north of the equator. Periodically, the initially straight front breaks and the two air masses start to form a spinning cyclone, rotating in an anti-clockwise sense.

Due to the fact that the earth is a rotating sphere, moving air experiences a pseudo-force which always acts perpendicularly to the air's movement, known as the Coriolis force. For air spinning in an anti-clockwise sense in the Northern Hemisphere, the Coriolis force always acts so as to pull it away from the centre of rotation. This leads to a dropping of pressure in the centre, which stops when a situation of balanced forces is reached – i.e. the pressure gradient pulling air in and the Coriolis force pulling it out are equal. This means that winds flow parallel to lines of equal pressure, normal to the line of maximum pressure gradient, despite speeds being defined by that gradient.

Once formed, the cyclones generally move North and East across the Atlantic, passing between Scotland and Iceland, but enveloping both. After about 8 days a typical depression will dissipate, only to be replaced by a new one coming in from the west. High wind speeds are therefore experienced on a regular basis in GB due to the transit of low pressure systems across it – particularly in winter. In the summer the cold and warm air meets further north, and consequently the low pressure systems form and travel further north, to some extent missing Britain. The eastwards movement of cyclones is guided by the jet stream, a narrow band of consistently high winds sitting above synoptic systems. The jet stream may deviate significantly from its more typical path, sometimes for a period of several months, giving rise to unusual weather patterns in GB.

Britain is also influenced by high-pressure systems, which often move in from the east. High pressure systems involve clockwise rotation and are larger than low-pressure systems, implying smaller pressure gradients, and gentle winds. They are known as anti-cyclones because the rotation is in the opposite sense of that of the Earth, while low pressure systems rotate in the same sense. High pressure systems bring clear skies, which mean low temperatures and therefore high electricity demand in winter. In summer they represent a convenient correspondence of low demand and fairly low wind power outputs. Rather than a binary choice of being dominated entirely by a low of high pressure system, the synoptic 'situation' over GB is in reality much more complex, defined by the presence of a combination of systems within an area larger than GB itself. There are several means by which this complex reality is categorised by meteorologists, and these are discussed at length section 2.3.1. It is noted by several authors, such as Oswald et al. [54], that the wind generation output in Britain can be almost nothing on winter days of very high demand. Such days are widely described

simply as high pressure situations, but as section 2.3.1 reveals, this is a gross oversimplification and the true meteorological situation is rather different.

Weather fronts are not the only driving factor for winds - in some parts of the word the daily pattern caused by the sun is dominant. It has been found in several studies that in Northern Europe there is a tendency for winds to start blowing more strongly in the morning and calm down in the evening, with the effect more pronounced during the summer [55]. The presence of diurnal patterns in GB wind generation is confirmed by Sturt & Strbac [56], while a consultation report prepared by NG [57] demonstrates a rather complex interaction of annual and diurnal seasonalities by means of a 'heat map' style plot of the average output of the existing GB wind generation fleet, with month of the year on one axis and hour of the day on the other. The plot confirms that diurnal variability is significant and greater in summer, but the annual seasonality accounts for more of the variance.

At a certain height above the earth surface, winds can be considered to experience no friction, and resulting flows (known as geostrophic winds), are determined by pressure gradients – along with considerations such as the conservation of mass, equations of state etc. The atmosphere below this height is known as the boundary layer, where winds are influenced by friction with the earth's surface and are often strongly affected by landscape features such as hills. The boundary layer extends up to two kilometres or more above the ground on a fine summer's day, but only about 100m on a clear night with low wind speeds. The difference is due to air close to the surface being heated and rising during the summer's day, i.e. much vertical mixing and therefore a downward transfer of horizontal momentum. However, on a cold night the air close to the surface cools more than the air above it, does not rise and there is little transfer of momentum.

The lower part of this layer is called the surface layer and is sometimes defined as a fixed fraction, e.g. 10%, of the boundary layer depth. When only considering meteorology relevant to wind power generation, one can often neglect the situations of lowest wind speeds and assume that the boundary layer extends to somewhere around one kilometre and the surface layer to about 100m - about the height of the largest modern turbines.

While wind velocities are 3-dimentional, this research is concerned only with scalar wind speeds in the horizontal plane since it is assumed that all generators of significant size have horizontal axes and may rotate within the plane so that they face the wind directly.

Within the surface layer the horizontal wind is well approximated as following a logarithmic law of increasing wind speed with height, as described in section 2.1.4.

2.1.2. The Spectrum of Wind Variability

Within the boundary layer there is always some chaotic mixing of horizontal layers, i.e. turbulence, which corresponds to relatively high frequency, broad spectrum and uncorrelated fluctuations in both horizontal and vertical wind speed. The spectrum of wind variability for the range 0.0007 – 900 cycles/ hour, i.e. periods of 60 days to 4 seconds, was investigated by van der Hoven [59]. He found that for vertical motion near the ground, the major contribution to the total variance is within a frequency range from 10 to 1000 cycles/hour (periods of 3.6 seconds to 6 minutes), i.e. turbulence. However in the case of horizontal motion, the variance within that range is only a small portion of the total.

Mainly examining one location (Brookhaven National Laboratory), van der Hoven found two major peaks in the spectrum, one at a period of about 4 days and a second at a period of about 1 minute. Between the two peaks, a broad spectral gap was found, centred at periods ranging from 6 minutes to 1 hour. Examination of data from other locations confirmed that this gap exists under varying terrain and synoptic conditions, albeit with some differences in range and point of minimum spectral intensity. A smaller peak is also present at a period of 24 hours, with a prominence dependent on the height at which wind is recorded. The peaks are interpreted as scales for which ‘eddy-energy’ may be well supported, and the lack of a physical process that can support eddy patterns of a certain size in the atmosphere is thought to be the reason for the spectral gap. The spectral density for the entire 0.0007 – 900 cycle/hour (4 seconds to 60 days) range at a single location is plotted in a figure generally referred to as the van Hoven Spectrum, a figure widely replicated in the wind resource literature. The spectrum is presented in figure 2.1 (the quality of the image in the electronically archived paper being quite poor).

Despite its popularity, cautious interpretation is required since the methodology behind the plot’s construction has serious flaws. These mainly relate to the fact that segments of the spectrum derived from very different and small datasets were simply joined together, with some ‘statistical corrections’. The periods for which higher frequency data were collected are short - hours to days, such that behaviour during only one specific type of weather was captured for each segment, and different weather types were present for the capture of

different segments. Additionally, different segments correspond to recordings at different heights above the ground.

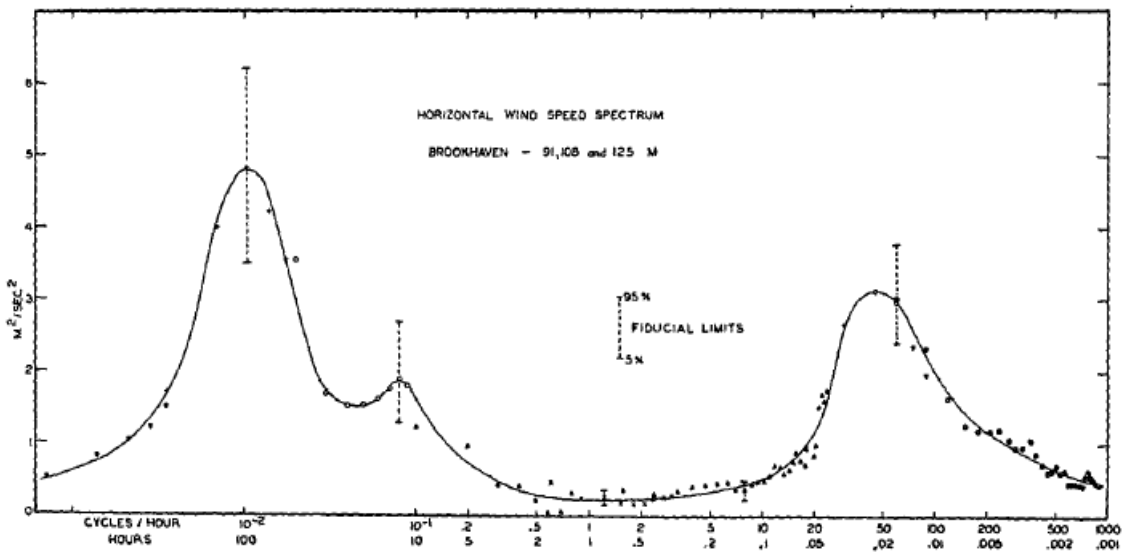

Fig. 2.1. Horizontal wind speed spectrum at Brookhaven National Laboratory at about 100m height. Extracted from [59].

Nonetheless, the methodology is probably sufficiently rigorous to allow the reader to conclude confidently that a clear spectral gap exists, separating variability associated with turbulence and lower frequency variability associated with low/ high pressure systems and seasonal patterns. In GB, a similar spectral peak to the one at 4 days might also be expected, corresponding to what Oswald et al. describe as the “periodic forming, moving, and dissipating nature of depressions” [54], although they suggest that the average period for the whole cycle may be closer to 8 days.

The implication of this gap is that the horizontal wind speed at an instant may be conceived as a sum of two parts, one with gradual temporal evolution and another which fluctuates rapidly, i.e. turbulence. By averaging over a period within the window, such as 1 hour, one eliminates the turbulence term and is left with a good approximation of the slowly evolving component. The wind resource during an hour can therefore be well represented by its mean speed, along with the turbulence’s variance, or variance divided by the mean speed. For the purpose of simplicity, this project assumes that the resource can be represented purely by the mean speed, ignoring the effect of turbulence on the hourly mean power generated implied by the highly nonlinear nature of the transformation.

2.1.3. Hourly Wind Speed Probability Distributions

Given a large sample (roughly ≥ 6 months) of hourly averaged wind speeds, their distribution is likely to be well approximated by the Weibull class of continuous probability distributions. For such distributions the probability density function is given by

$$f_W(x; k_W, C_W) = (k_W/C_W).(x/C_W)^{k_W-1}.exp(-(x/C_W)^{k_W}) \quad \text{for } x \geq 0,$$
$$= 0 \quad \text{for } x < 0, \qquad (2.1)$$

where $k_W > 0$ is the shape parameter, and $C_W > 0$ is the scale parameter of the distribution. The scale parameter C_W takes the value necessary to ensure that the function is a proper probability distribution, i.e. integrates to 1. It may be calculated using

$$C_W = \bar{u}/\,\Gamma(1 + 1/k_W) \qquad (2.2)$$

where $\Gamma(x)$ is the gamma function.

In the special case that $k_W = 2$, the distribution becomes the more familiar Rayleigh distribution, which arises when a random variable is the vector sum of two i.i.d. Gaussian variables. In other words if wind speeds in two orthogonal horizontal directions have identical normal distributions, then the resulting horizontal wind speed would have a Rayleigh distribution. It will never be the case that any location will have exactly the same distributions of speeds in orthogonal directions, due to prevailing wind directions and the influence of obstacles. However if they are similar and roughly Gaussian their vector sum will be approximated by a Weibull distribution with a shape parameter in the region of 2, i.e. the Rayleigh distribution. Such variables have positively skewed densities with a single peak, but for a shape parameter of $k_W \sim 3.6$, the distribution is close to the Gaussian. One would therefore expect a strong directional bias to raise k_W closer to 3.6.

Much of the standard datasets on wind speeds are recorded with cup anemometers, which typically have relatively high wind-speed thresholds, i.e. friction means that they will record zero wind speed even when this is not exactly the case. Histograms of wind speed measured in this way will therefore typically exaggerate the probability of calm or very low wind speeds. As a result, it is often more accurate to describe the probability density as a Dirac delta function plus the Weibull distribution, as noted by Tackle and Brown [60].

In addition to this false peak, distributions may contain several minor peaks, as discussed by Djokic, Matar & Hayes [61]. They propose that a mixed normal distribution, i.e. a distribution which is a weighted sum of normal distributions with different means and variances, is a better fit than Weibull. In this case, the number of normal distributions corresponds to the number of significant peaks in the empirical distribution. They found that 5

normal distributions gave the optimum balance between accuracy and computational burden for a couple of sites in GB.

2.1.4. Wind Shear

Wind speeds are usually recorded at a fairly low height, 10m in the case of the UK Meteorological Office, whereas modern wind turbines have hub heights of typically 70m – 100m (e.g. [62]), so it is essential to understand the relationship between speeds at the two heights. Unfortunately the complex and dynamic nature of the atmospheric boundary layer means that such extrapolation of wind speed from one height to another is always uncertain. It was stated in Section 2.1.1 that mean horizontal wind speeds increase according to a logarithmic law within the surface layer of the atmosphere. More precisely, the speed at height z within the surface layer is given by

$$u(z) = (u^*/k_{VK})\, ln(z/z_0 + \psi_S) \tag{2.3}$$

where u^* is known as a shear velocity, z_0 is the surface roughness length, k_{VK} is Von Kármán's constant and ψ_S is a stability correction. The roughness length parameterizes the roughness of the local surface and the shear velocity parameterizes the frictional force between the moving air and the ground. The shear velocity is in fact a way of re-writing shear stress τ and density ρ in units of velocity and is given by $u^* = \sqrt{(\tau/\rho)}$. The von Kármán constant is a dimensionless constant often used in turbulence modelling, typically taken to be 0.41. The stability term relates to the buoyancy of air – weather vertically displaced small volumes of air tend to keep on moving in the direction of displacement (unstable conditions) or rather tend to move back to their original position (stable conditions). For most wind energy applications ψ_S is assumed to be zero, corresponding to a condition known as neutral stability – which is often the case when wind speeds are high.

When the roughness length is small, i.e. the surface below the wind turbine is smooth, the wind shear gradient will be greater and winds at hub height will be stronger, as is well known from common experience. It is also well known that wind speeds are higher at certain locations such as the brow of a hill – in such cases the increased gradient is represented mathematically as in increase in u^*, arising from conservation principles for the flow.

While the logarithmic approach is based upon scientific principles, another methodology is commonly used by engineers, based on a power law. In this methodology, the wind speed u_2 at height z_2 is related to the speed u_1 at height z_1 according to

$$u_2/u_1 = (z_2/z_1)^{\alpha} \tag{2.4}$$

where α is the wind shear exponent which essentially amalgamates the stability correction, shear velocity and roughness length aspects of the log law into one factor. According to Kubik et al. [63] the power law is the method used in many of the most influential renewable variability studies and is also widely used by wind developers in site appraisals. Developers, they claim, tend to check using both laws, but as all the power law parameters can easily be derived from a wind mast, this is the preferred method. With the stability term ψ_S ignored, the two models have been shown to perform equivalently in shear extrapolation predictions, although at any particular site one model may be better than another. It must be noted that the coefficient derived from one height is not applicable to an extrapolation from another height.

Kubik et al. present a very important result: while a single annual average value for α may be selected to accurately represent the long term energy generation from a simulated wind farm, there are large differences between simulation and reality on an hourly power generation basis. Calculating true α values for the majority of hours in a 12 month period at an actual wind farm location in Scotland, the authors found that the coefficient was approximately normally distributed, with an annual mean of 0.119 and a standard deviation of 0.172. The fact that the standard deviation is greater than the mean, with some hours having negative values, demonstrates that assuming a constant value for α is a very poor approximation. The distribution has very heavy tails, with short lived extreme values reaching as high as 3.825 and as low as -3.648. No annual seasonality was found, but diurnal patterns were strong, with overall averaged nightly wind shears almost double that for daily values. The diurnal variation was stronger in summer, with mid-day values generally lower than winter mid-day values, and summer midnight values higher than winter ones. Since the diurnal pattern is in the opposite direction to that of 10m wind speeds, it seems that diurnal trends will in general be reduced in magnitude at hub height.

Wind speeds, and hence generation, tend to be overestimated at lower true wind speeds and underestimated at higher true wind speeds. Despite obvious trends, that is all they are, i.e. there is considerable apparent randomness present, with some of the lowest values occurring within a few hours of the highest. This study is extremely valuable in highlighting the complexity of this issue and demonstrates that if the current research project is successful in modelling the behaviour of the wind speed field across GB at 10m height, this is only part of the picture.

A different perspective is provided by an interesting plot found in Petersen et al. [52]. The plot shows the variation of Weibull parameters with height for a location in Denmark, described as having roughness class 2: 'agricultural land with some houses and sheltering hedgerows'. The plot shows that the shape parameter has a value of 1.8 at 10m, increasing to a maximum value of 2 at 100m and falling to a value of 1.75 at 1000m. The scale parameter increases from about 5 to 11 over the height range, following the log law to a good approximation. To summarise: the literature shows that the dynamics of wind shear are very complicated, it's study is a substantial and topic in its own right, and it is not possible to develop any accurate representation of it in the current work.

2.2. Previous Detailed Analyses of Wind Resources

This section reports on some of the relevant aspects of detailed wind resource studies, some of which are specific to the UK. In most studies discussed here it is wind power outputs rather than wind speeds examined, either metered outputs or generated from wind speeds via a (simple) generation model.

2.2.1. The Work of Dr. Graham Sinden

The section reports on the substantial and influential work of Dr Graham Sinden, which examined the wind resource in GB. The work is published in a report commissioned in 2005 by the UK Government's then Department of Trade and Industry [64], and in more technical detail in a journal article [65] (2007). Due to its thoroughness and relevance, it is reported in some depth here.

Sinden's methodology is to generate wind power outputs from historical hourly-averaged wind speed records collected by the UK Meteorological Office during the period 1970 – 2003. Sinden argues that analysing data from such a long period, as opposed to the much shorter periods for which power output datasets are available, provides confidence that the results will include low-frequency but high-magnitude climate events – similarly to the argument made in Chapter 1 that periods even longer than a few decades are needed. Such events include temporally and spatially extended high or low wind speed events that may not have occurred within the timeframe of wind power operations in the UK. Additionally, wind recording sites are located throughout the UK, while metered (i.e. transmission connected) wind farms were, at the time, all in Scotland. Sinden therefore chose to reflect the diversified wind power system that might be expected in the future, by taking data from 60 varied locations throughout the UK.

The historical datasets did not provide full records for every hour in the period. The methodology considered that for any hour, the presence of valid data in a minimum of 45 sites out of the 60 provides sufficient representation of spatial detail, so this was the criteria for accepting a particular hour for analysis. This meant that the majority of hours were included and the dataset was very large, and considered without the need for interpolation of missing hours.

The wind speeds at each location were converted into power outputs before calculating averages over all locations and expressing them as load factors - the percentage of

the maximum possible output that the distributed capacity can generate, for each hour. Sinden recognises that wind speeds at each location need to be up-scaled to reflect hub heights (assumed to be 80m) before being converted into powers. No mention is made of the log or power laws, but up-scaling is regionally specific and guided by categorisation into location types – 'coastal', 'inland' and 'island', as found in the European Wind Atlas [66]. The model for converting a set of wind speeds into load factors was validated with two independent tests. One was to compare the derived values for standard deviation of 1 hour-ahead and 4 hour-ahead changes in load factor for the model with historical values for the UK and Denmark. The other test was to compare the temporal pattern of annual capacity factors for the model and available historical data for the UK. The model's values compared well with historical ones in all cases - a strong indication of validity, but the model is based on constant up-scaling factors, which the previous section showed is not a valid assumption. Some caution is therefore necessary with regard to the details of results for specific hours, although conclusions involving averaging, upon which the report is somewhat too reliant, are probably accurate.

Sinden's analysis found that:

- A typical turbine will generate some electricity for 80-85% of all hours of the year;
- Wind generators typically operate below rated capacity for around 90% of all hours, (highly site dependent);
- Assuming a long term annual average capacity factor of 30 %, annually averaged load factors for the period 1970 - 2003 ranged from a minimum of 24.1% to a maximum of 35.7%, with a standard deviation of 7%. The biggest change between consecutive years was from about 34% to 24.1% in the years 1986–1987;
- The most changes in load factor from one hour to the next are less than +/- 2.5%. A change of about +/- 20% is likely to happen about once a year only; and
- On average, the wind resource can deliver about twice as much electricity in winter than in summer. Capacity factors are higher in the daytime than at night. This is most pronounced in summer, with overnight capacity factors of around 13%, and peak daytime capacity factors of around 31%. In summer the capacity factor is elevated for a longer period, 7am - 10pm, compared to 10am - 7pm in winter.

The spatial extent of low wind conditions is the subject of detailed analysis in the study. It was found that whilst low wind speed conditions can be extensive, there was not a single hour during the study period where hub-height wind speeds at every location were below 4m/s. On average there was only 1 hour per year when over 90% of the UK experienced

low wind speed conditions, and this occurs mostly in summer. Over the course of a year, low wind speed events affecting more than half of the UK are present for less than 10% of all hours. The frequency of extreme low wind events is highly seasonal, with e.g. events where 75% or more of the UK experiences low winds representing around 0.8% of all hours, but only 0.2% of hours in winter.

Results are captured in a histogram in the journal paper in which the % of the UK experiencing low wind was allocated to 2% bins. This provides a good idea for the validation of the model developed in the current project – the construction of such a histogram for 10m wind speeds for both the model output and the data on which the model was trained. The report's conclusions about the rarity of low wind speeds appear somewhat at odds with other reports, such as Oswald [54]. The explanation is probably that Sinden's model has greater spatial diversity, and that Oswald's [54] case study of low winds involved rare conditions for GB when the wind shear may be significantly less than its average value. Perhaps Sinden's definition of low wind speed should be extended to e.g. 6m/s, since turbines have a LF of only ~ 10% at this speed.

The spatial extent of high winds was also analysed, with 'high' meaning above turbine over cut-off, i.e. $\geq$ 25m/s at hub height. This high wind speed criterion is only an approximation of the decision rules regarding turbine shutdown at high wind speeds, as these rules often assess gust speeds as part of the operational decision making process. It is clear however that extensive high winds are rare, with no more than 43% of the country experiencing then at any given hour within the 34 year sample.

The correlations between 2080 pairs of onshore wind sites in the UK were calculated, and plotted as a function of the distance between the sites. It is presented as figure 2.2 below. The plot demonstrated that correlations generally follow a $exp(-d_p)$ pattern, where d_p is distance. The greatest correlation coefficient value is over 0.9 at about 50km, while the smallest is slightly negative at about 900km. Orographic features and differing orientations mean that many points deviate considerably from the best fit curve. For example two pairs have a correlation of about 0.24, but pair are separated by 900km, while the other only 230km. The $exp(-d_p)$ principle is nonetheless a useful insight, used in several subsequent models, as will be shown in Chapter 4.

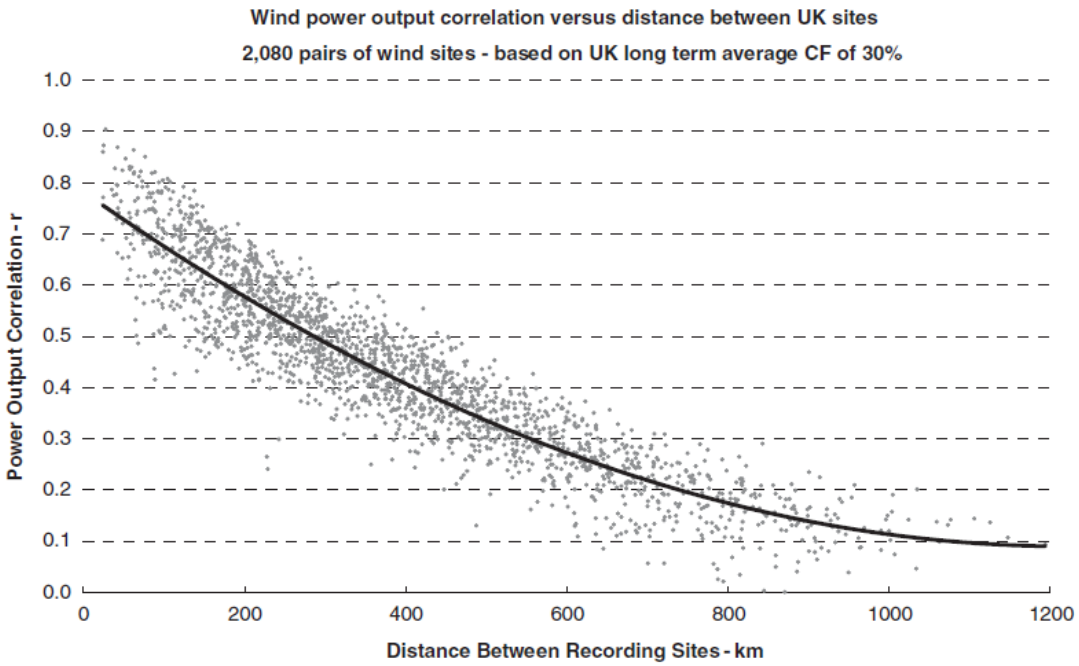

Fig. 2.2. Linear correlation coefficients for power outputs as a function of distance between locations, with a best-fit line included. Extracted from [65].

Sinden goes on to analyse the relationship between the UK's wind resource and demand. Hourly electricity demand data were obtained for England and Wales for the period 1996–2004, overlapping with the datasets to provide around eight years of combined data. The England and Wales demand accounts for 88% of total UK electricity demand, so the reasonable assumption made is that it may be scaled up to explore the relationship between resource availability and demand in the entire UK. Each hour of the eight year period was ranked in increasing order, and then grouped into one-percentile bands. The corresponding, i.e. time matched, mean load factor for each percentile was determined, and a considerable positive correlation found. The only exception is for the 92% - 100% range where the slope is negative, although this effect is fairly small and the 100th demand percentile has an above average wind load factor. The averaging is reported to hide considerable variability within each percentile – particularly for peak demands. This indicates that when simulating a power system with both wind output and demand represented, representing average historical behaviour would probably lead to inaccurate results, despite the need for simulated behaviour to converge to average historical behaviour at a suitable rate.

There is further analysis using the system of ranked percentiles, examining the maximum spatial extent of wind speeds below 4m/s occurring for each demand percentile.

The spatial extent is at its greatest for the lowest demand percentile, with the trend curve falling gradually from about 92% coverage to 75% at the 85th demand percentile, and rising again to about 81% at the 100th percentile. The relationship between demand and maximum extent of high winds is also explored and found to be quite complex. As a result, the relevant diagram has been included as figure 2.3 below. The trend line begins with a value of about 2% at the 0th percentile, rising to a peak of about 18% at about the 32nd percentile and with a smaller peak of about 13% at the 92nd percentile. There is a very big difference between maximum and average extents: a maximum extent of about 23% occurred during the 88th percentile, for example, but the average extent was a mere 0.2% (about 1 station) for the same percentile.

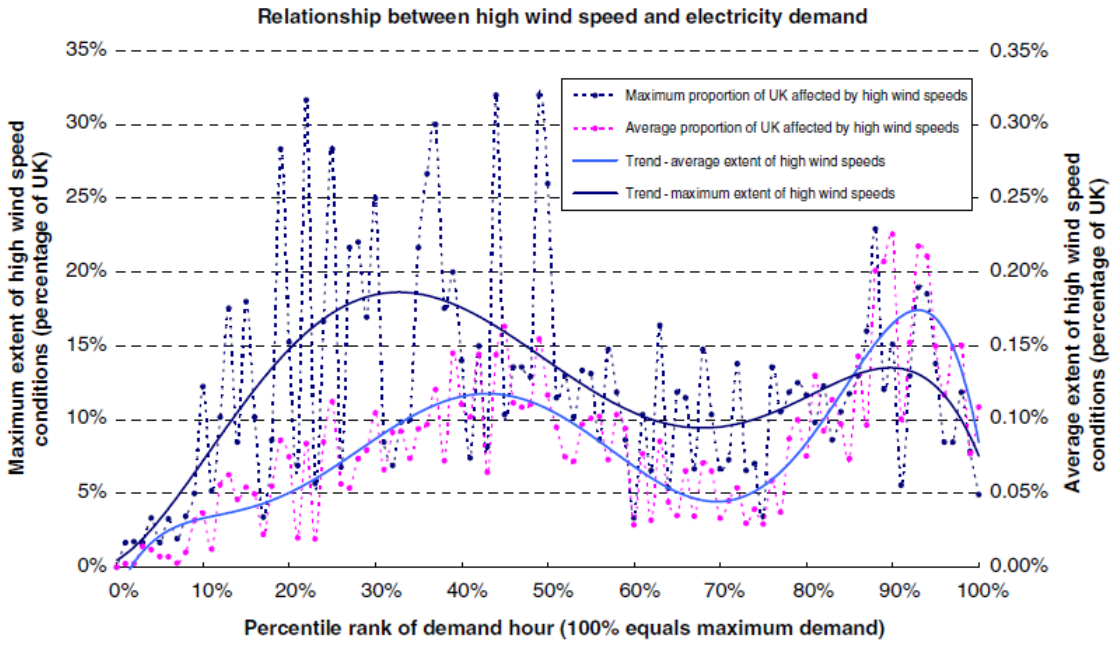

Fig. 2.3. The relationship between ranked electricity demand and both the average and maximum spatial extents of high wind speed across GB . Extracted from [65].

Other results presented in the paper involve two subsets of the total sample – one is peak demands, for which demand was in the 80th – 100th percentiles; and the other low demands, for which demand was in the 0th – 20th percentiles. The subsets were arranged and grouped by capacity factor percentiles and histograms are plotted, showing very distinct distribution shapes. While both subsets had occurrences of capacity factors in the full range from 0 to 100, the peak-demand subset was more evenly spread than the distribution for the entire set, i.e. less of a bias towards low capacity factors and probably a less positive skewness value. On the other hand, the low demand subset shows a much more pronounced clustering

at values below 10% capacity factor, and probably a higher skewness value. It is likely that similar differences exist between the distributions of 10m wind speeds for the two subsets.

2.2.2. Studies by NREL

The United States Government's National Renewable Energy Laboratory (NREL) and its subcontractor Electrotek Concepts collected wind power data, sampled at 1 Hz, from two large commercial wind power plants in the Midwestern region over several years. One wind farm was near Lake Benton in Minnesota and the other was at Storm Lake in Iowa. A paper was published describing the correlation between the outputs of these two large wind power plants [67].

The locations are about 200 km apart, and the terrain between them generally flat with little surface roughness. In this region, like GB, strong winds are generally associated with the movements of low-pressure systems. As a result, daily averaged outputs track each other very closely, with a correlation coefficient of 0.744 calculated for a 3 month recording period. On a monthly basis, the correlation coefficients of daily outputs are all positive numbers, ranging from 0.647 to 0.925. Examining values over two example years, a slight annual seasonality seems to be present, with summer months more likely to have higher correlation, but the pattern is highly stochastic. Examining data on a much finer temporal resolution – 1 minute averages, over a seven day period, revealed that correlations are at a maximum when a time shift is present, with Storm Lake leading Lake Benton by about 4 hours – representing the passage of a well-defined weather system moving from Lake Benton to Storm Lake over a period of 4 hours.

Not all correlation calculations yield positive values. On a daily basis, for example, the correlation between hourly average power of Lake Benton II and Storm Lake varies from 0.851 to -0.55. For one-minute average power data, the lack of consistent correlation is even more prominent, with values ranging from 0.942 to -0.987 within a single seven day period, with an average value of 0.054. Using 1-second power yields very similar results (0.920 to -0.961), suggesting very strongly that the high-frequency components of output power are nearly independent. This is good news in terms of the validity of representing the GB wind power resource with hourly speeds without turbulence, since independent high frequency fluctuations are almost entirely smoothed out by spatial diversity.

Despite the stochastic behaviour of their short-term power fluctuations, calculations of daily and monthly correlation coefficients with one-minute average power data provide

further evidence that, in longer time frames, the power outputs are highly correlated. On a monthly basis, the correlation coefficients are remarkably high and consistent, ranging from 0.530 to 0.757. It is clearly a requisite for the model developed in the current research to generate datasets in which the cross-correlations of hourly averaged wind speeds at different locations display a decreasing variance as sample sizes increase, tending towards the correct long term values from the training data.

NREL also produced a technical report [68], which again uses 1 Hz resolution power data collected over several years, at the same two locations and 3 more: one relatively close and two further, in Texas. Wind speed data was also obtained for 1 location in the Mid-West, at hub height and with one minute resolution. The report continues the focus on cross-correlation at different time scales, along with analysis of rates of change in power output, known as ramping rates.

A plot of simultaneous wind speeds and power outputs shows that power output follows the general trend of wind speed closely, but appears to be smoother, with the exception of a few steeper changes that are quite rare. This is due to both the nature of turbine power curves and that instantaneous outputs from individual turbines of a large wind power plant are not synchronized, despite their close proximity. The modest physical separations and differences in terrain details cause wind speeds at each turbine to vary, making the aggregate output from large numbers of turbines less volatile, as is the case for two separate wind-farms. This further validates the approach of ignoring the effects of turbulence.

The authors point out that simply multiplying the output of one turbine or a group of several turbines to find the total output of a farm is not, therefore, a valid methodology as it makes fluctuation characteristics worse than actual performance. The study confirms that during normal operation, the short-term fluctuations of wind farms are very small: about 0.1% of plant capacity for 1 second averaged wind speeds, less than 1% for 1 minute averages, and from 3% to 7% for hourly data series. Clearly, wind speeds do not change suddenly over a wide area to affect every wind turbine in a large wind power plant at the same time.

The operations of wind power plants in different regions and with different types of turbines were found to be very similar. The output power variations are a function of short-term wind speed variations, which were found to be similar everywhere. However, different regions have different daily, monthly, and seasonal wind power profiles. The report concludes

that wind power data from one region can be used in other regions to predict fluctuation behaviour from second-to-second up to hourly, but they are not good indicators of daily and monthly performance. This validates the initial assumption in the current project that the model developed must be trained on site-specific data for each location to be represented.

2.2.3 A Top-Down Approach

A cautionary note is provided by de Castro et al. in [69]. They note that almost all analysis work on the potential to generate energy from wind in the future is 'bottom-up' in nature, extrapolating from current wind speeds or power outputs to model, in some cases, very large capacities. This approach, they argue, fails to consider the atmosphere's energetic balance, potentially violating energy conservation.

An alternative top–down approach is developed by Castro et al. to evaluate the global technological wind power potential, while acknowledging energy conservation. Their results give roughly 1 TW as a top limit on the future electrical potential of wind energy - a value much lower than previous estimates and even lower than some published estimates for the economic and realizable potentials by the mid-21st Century. Their insight is certainly worth noting, but no practical means exist for integrating this limit into the model developed in the current research.

2.3. Synoptic Classification

2.3.1. Context

It was stated in section 2.1.1 that the synoptic situation over GB cannot be adequately characterised simply as having either a low or high pressure system directly over it. Rather, wind speeds and other weather variables such as temperature may be determined by the presence of several weather systems in an extended area which includes continental Europe, Scandinavia and the north east Atlantic. More accurately, wind patterns in GB are characterised by the growth and decay of synoptic-scale systems within this extended area, each lasting from one to several days and taking a variety of pathways.

It is useful to divide synoptic-scale situations into a set of characteristic types – a process known as synoptic classification, with a long history in meteorology and climatology. Surface weather based classification systems have been invented and implemented, each typically defining a fixed set of weather patterns, or criteria associated with such patterns, and then classifying the current synoptic situation according to this set. In the early days of meteorology, classifications were usually called catalogues of synoptic types and were used mainly in weather forecasting. James [70] states that although there may be markedly different weather in different regions of the extended area, interrelated large-scale characteristics can usually be determined.

When the weather pattern persists in a similar configuration for a number of days, the event is referred to as a weather regime. On some occasions, weather regimes may persist for a few weeks, in which individual synoptic systems follow very similar paths to their predecessors time and time again. Such persistent regimes typically result in locally extreme anomalies of weather parameters such as low temperatures and heavy snowfall – drivers of peak demands.

For these reasons, climatologists prefer to examine persistence in the large-scale atmospheric circulations responsible for the weather conditions and use this as the basis for classification. The circulation may be represented by either, or both, surface pressure fields or geopotential height fields: the height above mean sea level at which pressure has fallen to a chosen constant value, after adjusting for changes in gravitational field strength at different latitudes. The main applications of circulation classifications are in historical climatology, in analyses of recent climate variations, and in analyses of outputs from global climate models. A classification system may also be a very useful tool for assessing whether a long synthetic

series displays similar patterns of average behaviour, and similar occasional deviations from 'normal' patterns, as are seen in a long historical series. They are also of interest as possible candidates for defining discrete hidden states in a hierarchical Markov or semi-Markov model.

A considerable variety of classification systems exist for the atmospheric circulation over Europe. Cahynov and Huth [71], for example, analysed 23 systems. Of these, the synoptic class for particular days are determined on an objective basis in 18 of the systems, while selection is a subjective process for the other 5. The methods used in the objective classification systems analysed by Cahynov and Huth include, among others: (i) various kinds of cluster analysis, from k-means to simulated annealing; (ii) correlation-based methods; (iii) principal component analysis; and (iv) neural networks, namely self-organizing maps. Principle Component Analysis (PCA) is a method of identifying independent modes of temporal variability in multivariate processes, which may be common to any number of the variables to differing extents, and sorts them in decreasing order of contribution to total variance in the process. Huth et al. state in [72] that cluster analysis is frequently preceded by PCA to remove colinearity from the input variables that may negatively affect results of a clustering procedure, mainly by giving excessive weight to strongly correlated variables.

James [70] states that objective methods such as cluster analysis yield spatial patterns which, despite representing the most dominant modes of the variability, are too large-scale and smooth. They do not, according to James, intrinsically capture real synoptic characteristics. Additionally, some rare but nevertheless significant synoptic types do not appear in the output. There are therefore advantages of using (subjective) synoptic experience to pre-select a set of basis patterns manually, for subsequent classification. A well-known manual system, appropriate for North West Europe, is Lamb's daily weather types [72], in which the basic flow direction and level of 'anticyclonicity' or 'cyclonicity' over the British Isles is determined on a daily basis. The focus exclusively on a fairly small region may be seen as disadvantageous in terms of insight provided, and the German Grosswetterlagen (GWL) system is considered by many authors to be conceptually superior [70].

The GWL system was developed during the latter half of the 20th Century, and captures the large-scale characteristics of weather regimes while still focussing on local detail. Additionally, a catalogue of GWL types created by Hess & Brezowsky, going back to 1881 [70], is the longest European circulation classification record available. This system is the subject of the next section.

2.3.2. The Grosswetterlagen Classification System

The GWL system of classification is based on the division of atmospheric circulation over Europe and the North-East Atlantic into 29 types [70]. Developed in Germany, the system's primary focus is on central Europe and the synoptic type names (in German) relate to the experience of weather spells in this region. Examples of the synoptic types, referred to as GWLs, are: 'Anticyclonic South-Westerly', 'Icelandic High, Trough over Central Europe' and 'High over the British Isles'. The GWLs last at least 3 days, by definition, and any transient patterns are classified either to belong within a long-lasting GWL type or to be the result of a transitory phase between two GWL types. A table in [70] lists mean event length, averaged over all occurrences of each type in a 46 year period. There is surprisingly little variation across the types, varying from between about 4.5 days for some rare and dynamic types, and nearly 6 days for two westerly types.

As stated above, a daily catalogue of subjectively-assessed GWLs for the greater European area was constructed by Hess and Brezowsky (HB-GWL), retrospectively going back to 1881 and extended to the present by the German Weather Service [70]. In order to overcome problems of subjectivity and an excessive focus on central Europe, James [70] presents an objective methodology for the GWL. His methodology involves production of composite base patterns for each GWL from the HB-GWL catalogue and then employs pattern correlation calculations to generate an objective catalogue. Separate composites are constructed for summer and winter, and correlations are calculated for appropriately time varying superpositions of them. Fairly complicated logical filtering is required to satisfy the condition that each type must be at least 3 days in length. The objective-GWL classification is based upon both fields mentioned in section 2.3.1: mean-sea-level pressures (MSLP) and Geopotential Heights at 500 hPa. The latter quantity is the height above sea level at which atmospheric pressure has dropped to 500 hPa, with corrections made so as to cancel out the effect of latitude on the Earth's gravitational field strength (since it is not a perfect sphere). These are calculated as daily mean fields at a 1^{o} x 1^{o} resolution from a re-analysis dataset, ECMWF ERA40 [73], covering the period September 1957 to August 2002.

Meteorological re-analysis datasets are the outputs of data assimilation projects, which assimilate historical observational data spanning an extended period, using a single consistent assimilation scheme. The process of data assimilation involves analysis of the atmosphere's state for each hour, ensuring consistency – i.e. correction of measurement

biases and erroneous entries, and the filling-in of missing data based of numerical model predictions.

A table in [70] presents the total number of occurrences of each GWL type for each month of the year during the re-analysis period, revealing moderate but complex annual seasonalities in their occurrences. Another gives the number of day-occurrences of each type separately for each year in the period, revealing that occurrences of relatively rare types are sometimes strongly clustered over a periods of e.g. 6 years. A transition matrix between types is presented, which is far from sparse. Transition rates are generally quite spread out, although the GWL type probability density for the 'new' state is noticably conditional on the 'old' state.

A table is also included of the longest consecutive number of days recorded for each type, as well as the longest recorded absences - showing that even the most common GWLs undergo phases of conspicuous absence. For example, the common westerly type WZ was absent for nearly a year in 1959 - a year noted for a long, dry and warm summer over north-west Europe. In an example of clustering, another similar and common westerly type, WA, remained absent throughout the following year, 1960. Even the rarest GWLs have occurred as continuous sequences of at least 10 days at some point, while some remarkably persistent spells of common progressive types lasting more than 3 weeks have been recorded. James states that during such long sequences, occasional transients will have disturbed the circulation, but not sufficiently to have broken the sequence. In other cases, a slightly stronger or more persistent disturbance may indeed 'officially' break a sequence, reflecting an unavoidable arbitrary aspect of any classification system.

One of the main proposed applications for the wind datasets to be developed is establishing wind generation's contribution to meeting extreme peaks in electricity demand for future scenarios. As discussed in Chapter 1, robustly assessing this contribution directly from power system data and meteorological records is difficult since extreme peaks occur infrequently (by definition) and measurement records are short, noisy and inhomogeneous. Zachary, Dent and Brayshaw in [31] and [74] propose atmospheric circulation-typing combined with meteorological reanalysis data as a potential means of addressing some of these difficulties. Where statistical data is very limited, they argue, it is necessary to use physical insight about the problem to derive robust conclusions. In the case of wind and demand, this means analysing the weather systems which have historically driven extreme demands.

The “High-over-Britain” anticyclone, of concern to authors such as Oswald et al., is found to be generally associated with very low winds but relatively moderate temperatures, and therefore moderate daily peak demands. The coldest temperatures, and therefore highest demands, are rather associated with longitudinally-extended ‘blocking’ high pressure systems over Scotland to Scandinavia and latitudinally-extended low pressure ‘troughs’ over Western Europe. In both situations, wind-resource averaged across GB appears to be moderate.

Zachary, Dent & Brayshaw found in [74] that during the period from 1986-2005, the top two percentiles for hourly demands in England and Wales were grouped into only 15 separate days. Of those days, 10 had ‘blocking’ GWL types (7 HNFZ, 1 HNA, 1 SEA and 1 HFA). Of the remaining 5 days, only one was a ‘High over Britain’. Examining demand across GB during the winters 2001/2002 – 2009/2010, they discovered that all hours in the top demand percentile for GB demand for occurred on only 5 occasions, ranging from 1 to 8 days in length. There is again a dominance of blocking types, along with one trough and two ‘others’. The longest period was in 2010, a year for which the wind resource was exceptionally poor [75].

A scatter plot is presented in [74] of 10m height mean wind speeds and 2m height mean temperatures for each GWL type during winter. The mean values were obtained by averaging all the daily-mean wind-speeds and surface air temperatures that occurred during a given GWL type between November and March in the ERA-40 re-analysis dataset [73]. Both involved spatial averaging over GB, more precisely the area bounded by the ranges 8^{o}E - 3^{o}W and 49^{o}N - 59^{o}N. The mean wind speeds range from 3 to 8.5^{o}C for the types, while wind speeds range from 5 – 8.5m/s, and a clear positive correlation exists between the two variables. Blocking types are clustered closely together at the low temperature edge of the plot.

In addition to changes in mean temperature and wind-speed, Brayshaw, Dent and Zachary comment in [74] that the differing dynamical properties of the high-demand circulation types (High over Britain, blocks and troughs) suggest that the statistical link between electricity demand and wind availability may be rather different in each case. For example, troughs are dynamically active storm systems, moving rapidly eastward and commonly associated with both cloud and precipitation. Blocks on the other hand can persist for longer, and generally bring drier air and stable conditions over Britain. These two types of systems will therefore have very different impacts upon the properties to which energy demand and wind-supply is sensitive, such as ambient light and more importantly here, wind-shear. This mirrors Sinden’s observation in [64] that there is considerable variability around mean statistical relationships at peak demand times.

During an unpublished seminar presentation to peers [76], Dr Brayshaw presented slides demonstrating that each high-demand circulation type has a considerable spread of points in wind-temperature space, for daily averaged values. Brayshaw fitted bivariate Gaussian distributions to the points in each weather type – despite a scarcity of available points for the rarer types. Subsequent cumulative probability contours are all ellipses, some have a significant 'tilt' reflecting a correlation between wind and temperature. There is considerable overlap between types, implying that knowing the spatially averaged wind speed across GB provides little knowledge of the synoptic type. This does not indicate whether knowing the *pattern* of wind speeds across GB is a better differentiator of synoptic type – that is a question worth investigating in the current project, probably via PCA. Intuitively, it seems likely that the lack of wind direction information inevitably implies uncertainty.

Concerned that 29 might be a higher than optimal number of types for a single region such as GB, Brayshaw and Masato [77] have reduced them to 7 and 6 types for winter and summer, respectively, using a hierarchical clustering algorithm. Following the spatial distribution of the composite pressure fields, the new winter types (referred to in the journal paper as regimes) were named: Westerly [W], Atlantic ridge [ATLR], European trough [EURT], Atlantic trough [ATLT], Atlantic blocking [ATLB], European blocking [EURB] and European ridge [EURR]. The 6 regimes for the summer period are similar to the winter ones, with the first 5 of them identifiable as summer counterparts of W, ATLR, EURT, ATLB and EURB, while the last may be seen as a mixture of the winter regimes ATLT and EURR. Using the new reduced system, the 15 extreme demand events discussed above correspond to 8 ATLB events, 5 EURR and 2 EURB.

With the number of types reduced in this manner, any conclusions about differences in their internal dynamics will be much more robust, and they seem much more plausible as candidates for the regimes in a hierarchical semi-Markov model. If it is decided rather that a regression or straightforward Markov model is the best choice, there remains value in exploring the possibility of associating each day in the generated synthetic datasets with one of the concatenated GWL types – both for validation purposes and to correlate wind availability and demand more realistically in Monte Carlo simulations.

2.3.3. Wind Field Clustering

Despite the overall superiority of large-scale atmospheric circulation classification systems where types are based on meteorological insight, there remains value in schemes

based directly on objective classification of surface wind measurements. The most obvious advantage of the latter in the current context is that accurate identification of the correct class for the synthetic data is undoubtedly possible. Several authors have in fact established cases where cluster analysis of the wind field can determine a set of recurring meteorological states, without any a priori knowledge. Unfortunately, the clusters are generally based on wind speeds in orthogonal directions. The main application to date for such systems is in the regional-scale modelling of air quality, as demonstrated by Beaver and Palazoglu [78], and Glascoe et al. [79] – both of which refer to many similar studies.

Working with hourly averaged data from a series of recording stations in the San Francisco Bay area, Beaver and Palazoglu employ a sophisticated methodology designed to ensure that variance unrelated to synoptic scale patterns, which they identify as being on diurnal and annual scales, does not introduce periodic biases into their cluster labels. As is the case for large-scale circulations, the sequence of cluster membership should not be expected to change rapidly, the authors state, but rather should be comprised purely of multi-day periods. Short periods of rapidly changing cluster membership should be considered as either outliers or unstable transitions between stable atmospheric states.

The 4 wind field classes identified by Beaver and Palazoglu are found to correspond to larger scale atmospheric circulations. One cluster, for example, captures a high pressure system over the western United States where anti-cyclonic winds block the typical marine flow through the study region. Another pattern is described as representing a seasonal, offshore ridge of high pressure which reduces marine flow and produces a shallow boundary layer.

Five classes are identified as the optimal number by Glasgoe et al. in [79], who do not identify the specific region of the United States to which their results apply. They examine the extent to which surface winds are influenced by large-scale circulation – finding that July, for example, is characterised by weak synoptic forcing and local wind patterns dominant. The fact that 4 and 5 classes, respectively, are identified in the studies seems consistent with the observation reported in section 2.1.3 that a mixed normal distribution, with 5 peaks, was a good fit for the distribution of horizontal wind speed at a studied location. In other words, it appears that wind may be characterised as being in one of a small number of states, each with their own normal distributions.

On a somewhat tangential note, it is worth mentioning a methodological detail included in Beaver and Palazoglu, regarding the filling-in of gaps in the recorded datasets.

Short gaps of one or two hours for a particular variable, they state, can be interpolated across time. However, larger gaps in the series are filled-in using a method that assumes a multivariate normal distribution among the variables and fills the missing values using ‘expectation maximization’ – which surely means calculation of conditional means. Although this method does not consider the chronology of the data, it is said to perform well for the case when observations are only missing from a few stations simultaneously.

2.4. Climatological Variability

2.4.1. Introduction

Variability is an intrinsic feature of climate, with weather patterns changing not only from year to year but also between consecutive decades. The data upon which wind speed models are based necessarily cover a limited period of time only, and Chapter 4 will show that this has traditionally been a few years to a decade, certainly no more than two decades. The impact of long-term climate variability on wind power output patterns has been largely neglected to date, and not much attention has been paid to the questions:

- To what extent is the training data period representative of the longer-term climate? And
- How large should deviations between years and decades be in the model output?

This section will explore such questions. In this context, 'climatological' time scales mean a season or more in length.

Petersen et al. [52] report on the results of a study showing that in Northern Europe, variations in wind energy yield of up to 30% can be expected from one decade to another, while another study shows that expected annually averaged power outputs for a 45-m high wind turbine have a relative standard deviation of approximately 13%. A very thorough investigation of low-frequency variability was conducted by Palutikof, Kelly and Davies [80], its rigour and relevance justifying detailed presentation of results below.

The analysis is based on two datasets, one they describe as a relatively short-term time series of monthly averaged 10m wind speeds for a set of 52 meteorological stations, dispersed through the UK (49 of them in GB), for the period 1956-82. The authors also compiled a set of 6 longer time series, again of monthly averages, from 1898-1957. The results for one of the series is presented – Southport, on the west coast of England. The site is 7m above mean sea level, with an open and flat exposure in all directions. The dataset is considered homogenous since the same instrumentation was used for the entire period.

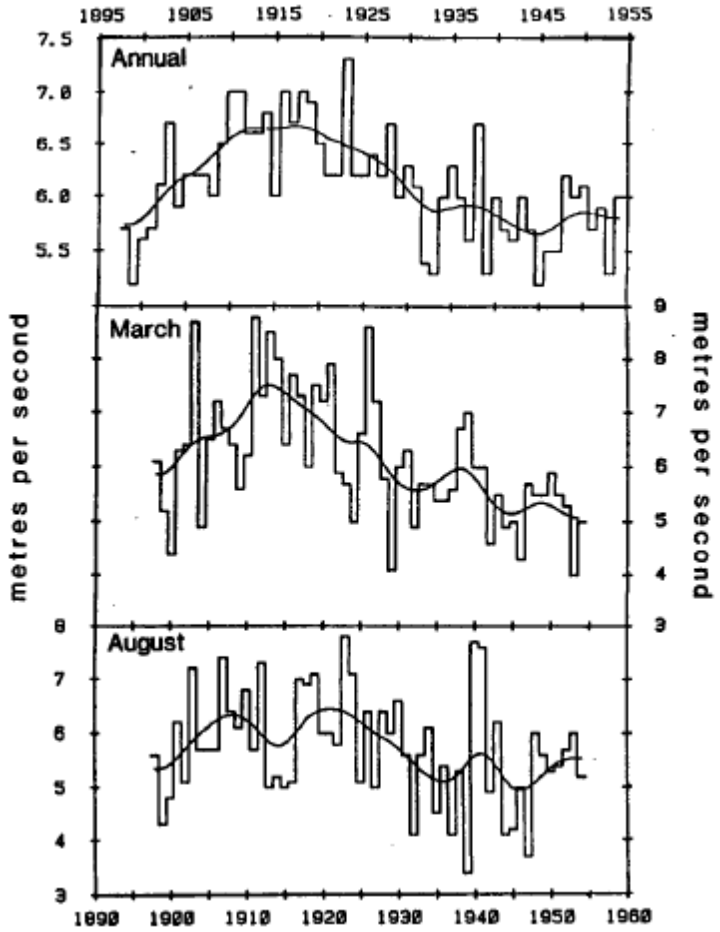

Fig. 2.4. Annual, March and August mean wind speeds for Southport (England), 1898-1954. Extracted from [80].

Annual averages were calculated and smoothed with a 15-term Gaussian filter, and their plotted results are presented in figure 2.3. This smoothed curve shows that at the beginning of the 20th century, average wind speeds were about 5.7m/s but that they increased rapidly to 6.5m/s in the 1910s. They fell back to under 6m/s in the early 30s, and remained at roughly that level for the remainder of the sample period. Such smoothed series were produced for each month separately, revealing that it is the variability for winter months that mainly drive the patterns in annual means. Over the complete time series, the maximum annual mean wind speed is 7.3m/s, in 1923, and the minimum is 5.2m/s, in both 1899 and 1945. Based on a series of reasonable assumptions, such as that shorter term wind speeds have a Raleigh distribution, the authors calculated that this range of average wind speeds represents a change in the average load factor of from 33% to 16%. The paper's results for the shorter but more spatially detailed time series are presented in the next section.

2.4.2. Low Frequency Variability and Large-Scale Circulation.

Having developed an understanding of the dynamics of large-scale atmospheric circulation types in the section 2.3, we now examine how low-frequency climatological change is related to their occurrence. This is achieved initially through continued reporting on the work of Palutikof et al. [80], specifically their spatiotemporal analysis of UK wind speeds, using the more spatially detailed dataset mentioned in the previous sub-section. As a short aside: the report states that wind speeds are lowest over inland southern England, increasing towards the north-west of GB, with the long –term annual wind speed in the Outer Hebrides over 7m/s. They also note that variance displays a similar spatial pattern, and that the time series for each location display different characteristics. Despite differences, sites are tied together through their relationship to large scale circulations, and these relationships are explored by means of PCA.

Having eliminated annual seasonalities by expressing the monthly wind speeds as anomalies from the long-term monthly mean, it was found that the 1st principal component (PC1) had very similar loadings (relative weight of contribution) from each station, bar 3 in the extreme north. This, combined with the fact that this component accounts for 51% of variance in the dataset, implies that PC1 represents well the large-scale characteristics of the wind field over GB.

The variation of each component from month to month is expressed as a time series of dimensionless amplitudes, known as scores. The PC1 scores were aggregated into yearly scores and plotted. Trends lasting several years were found in the series – all values being positive from 1962 to 1967, for example, while the decade 1968-77 is dominated by negative scores. When aggregated separately into 6-month summer and winter averages, it was found that during the 1968-77 period, values were generally negative for both seasons, such that they act in concert to produce negative annual means. From 1978 onwards, annual means are generally close to zero, as a result of considerable scatter in winter and summer scores. This means that there exist extended periods of several years where wind speeds across the country are consistently lower or higher than their seasonal averages, while other periods exist where there is no very-long-term pattern and seasons may fluctuate between having below or above average wind speeds.

Crucially, the authors also examine the relationship between such low-frequency variations and large-scale circulation types. They choose to use the Lamb index of circulation over GB, mentioned in section 2.3.1, which allows 27 different circulation types. They refer to

another study in which PCA was performed on a dataset of annual frequencies of the 27 types, from 1861-1980, that found periodicities on similar time scales to the wind field. The PC1 for the Lamb types accounts for 34% of total variance and contains significant loadings from only two basic weather type: anticyclonic (positive loadings) and westerlies (negative loadings). A comparison of PC1 scores for the two PCAs, for the 19 overlapping years, shows a very strong correlation, with a correlation coefficient of -0.902 (arbitrarily negative because of the loading signs in the Lamb PCA). In other words, generally high wind speeds over GB are, unsurprisingly, associated with a predominance of westerly days at the expense of anticyclonic days, and vice versa. What is most interesting is the low frequency patterns in the relative frequency of occurrence of the important circulation types, which defy easy characterisation.

Palutikov et al. were also interested in whether wind speeds in all areas of GB are equally affected by large-scale circulation patterns. An index of relative frequencies was constructed, based on the frequency of occurrence of anticyclonic (A) and westerly (W) type days. For a month where westerlies dominate, the index is negative, as was the case for the PC1. Individual time series of the A/W index were compiled for each month of the year, i.e. we have 12 series, each with 19 terms. For each of the 12 months, correlation coefficients were calculated between the A/W index and the mean wind speed series for that month, for each of the 52 stations. The number of months per year for which a statistically significant cross correlation exists was calculated for each station. It was found that 41 of the 52 stations in the dataset have six or more months with significant correlations.

A map of the UK is presented with contours for the number of months for which there is significant correlation. North of the latitude of central England the contours follow a reasonable approximation to straight lines in a North-North-West direction. The dependence of wind speed on the frequency of westerlies therefore increases gradually from the East coast to the West, more or less, as intuition would suggest. In southern England however the contours are much more convoluted. For example, a site in the Thames Estuary has only 3 correlated months while another about 65km to the North-East of it has 10, entirely contrary to the usual pattern. The lowest number of correlated months for a site is 1, in the far north east corner of Scotland, while the highest number is 12, for RAF Valley on the Isle of Anglesey.

A topic closely related to the dynamics of annual and monthly averaged wind speeds is regional wind indices, as discussed by Harman and Morgan [81]. Wind indices are described as historical time series of wind energy production for a geographical region, normalised to a long term average and usually produced on a monthly and/or annual basis. They are also

(presumably) normalised by installed capacity within the region, and are therefore statistical tools providing long-term trends for the region. They should ideally be based on actual wind farm production figures, and such series are publically available for the Danish, German, Netherlands and Swedish markets. Unfortunately, such a series is not produced for the UK, as is true of several other active markets including Spain and Ireland. Harman and Morgan provide a plot of the monthly and annual series for Denmark between 1992 and 2004. A clear annual seasonality can be seen, but the index values for peak production months for individual years vary from about 120% for a poor resource year to just over 200% for a good one. The minimum production months vary from about 20% to 75% for individual years, while the annual averages vary from about 80% to 110%.

For northern European countries for which there are no indices available, the authors used their position as employees at wind consultancy Garrad Hassan & Partners Ltd (GH) to create them for themselves, based on wind speed measurements. Plotting the real and modelled annual indices together for Denmark, Germany, UK, Ireland, Sweden and Holland, for 1993 – 2004, reveals that they are all highly correlated, particularly high for the UK and Ireland. They go on to plot the average index for the set of countries, for the same period, along with a measure that receives considerable attention for the remainder of this section: the North Atlantic Oscillation Index (NAOI). The NAOI is a measure of pressure difference between Iceland, where low pressure systems typically reside, and the Azores in the middle of the Atlantic, where there is usually a high pressure system. It is an average value for this pressure difference during a season or year, and is a key tool used by Meteorologists to understand long term weather cycles across Europe. Harman and Morgan express NAOI values as percentages of their long-term average, but it is more common to subtract this average before dividing by it. As such, when the NAOI is greater than its long term average value, the period is described as NAO+, while it is NAO- if the opposite is true.

The plot presented by Harman and Morgan proves the relevance of the NAOI, since there is a very high degree of positive correlation between it and the spatially averaged north European wind generation indices. The only caveat is that when the NAO index dips to an extreme low value of about 67% during the analysed period, the wind index only falls to about 92%. A season in which the NAOI is significantly positive or negative represents a persistent atmospheric circulation pattern, or circulation regime. According to Huth et al. [72], the theory of circulation regimes is originally based on the chaos theory where attractors (the regimes) pull the system of circulation states onto preferred paths around them.

In view of a strong increase in the industrial exploitation of the North Sea, Sušelj, Sood and Heinemann [82] address some questions related to the climatology of the daily mean wind speed at 10-m height in this region. Areas of investigation included the relationship between the monthly mean wind speed field and the large-scale circulation patterns over the European–Atlantic region, and their inter-annual variability in the last few decades. They confirm that the NAOI is well correlated with the wind speed over northern Europe, and that it is stronger in the winter season where the large-scale circulation has a more dominant influence on the weather.

The article's authors investigated whether a regional spatial pattern of Sea Level Pressure (SLP) variability exists which has a comparable or more dominant influence on the wind speed variability over northern Europe than the NAO. They conclude that most of the inter-annual variability of the wind speed field is explained by the first component of a PCA, known as an Empirical Orthogonal Function (EOF) in this case. The principal EOF pattern describes a coherent increase of wind speed over the North Sea region and is related to a SLP pattern similar to the NAO. The northern centre of the pattern found in their study is shifted eastwards, however, similar to the dominant pattern found by some other researchers.

The second SLP pattern is the anomaly between Scandinavia and Greenland and is related to a wind speed dipole over south and north of GB. The index from the second circulation pattern was found to be well related to the wind speed field in the recent period, whereas it was rather poor in the past period. The authors do not know if this represents a genuine change or is a reflection of better data quality. The second SLP and wind speed patterns is seen as reflecting the monthly mean position of the storm tracks and increases toward the end of the reanalysis period, which is consistent with an observed northern shift of storms tracks in the recent period (although Brayshaw et al. in [83] relate the movement of this track to the NAOI). This work therefore confirms how intimately related seasonally averaged wind speeds are to the NAOI, but that this is not the only mode of variability which influences wind patterns. Those patterns were found to be subject to changes which either are not cyclical or are cyclical on a scale much longer than a decade.

Commenting that research into wind speed variability on climatological time-scales is in its infancy, Brayshaw et al. [83] examine the dynamics of a modelled wind turbine under three different predefined states of the NAO during an extended winter season, November to March. In the previous example there were 2 states of the NAO: the NAO+ state corresponding to NAOI values above the long-term average, and NAO- corresponding to NAOI values below

that average. In Brayshaw et al.'s model, the three states are "high NAOI" (NAOI $\geq$ +0.5), "medium NAOI" (+0.5 > NAOI > -0.5) or "low NAOI" (NAOI $\leq$ -0.5). Results confirm that the NAO has a statistically significant impact on the hourly, daily and monthly-mean power output distributions from the turbine, although these differences are not large.

In each NAOI state (as defined above), the wind speed distribution resembles the Weibull distribution, although wind speeds are generally higher for the high NAOI state, and more clustered around a single, fairly low value for the low NAOI. Brayshaw et al. examined the impact of the NAO state on power outputs extrapolated from hourly averaged wind speeds at an exposed site in the North West of England (Great Dun Fell). It was found that in the high NAO state the rated power is achieved in more than 40% hours, compared to the low NAOI state, where the rated power is achieved in approximately 30% of hours. For both Great Dun Fell and Stornoway, in the Outer Hebrides, there is a roughly 10-15% difference between the mean power output estimate in the high and low NAOI states. (No summary values were provided for the 'medium NAOI' state).

The effect of the NAO state becomes much more pronounced and complex for future electricity system scenarios in which there are (i) higher capacities of different renewable generation types, particularly hydro-power, and (ii) greater reliance of interconnectors between countries/ regions such as GB and Scandinavia,. This is due to the fact that the NAO affects mean levels of wind, temperature and precipitation across northern Europe, and was explored by Ely [84]. While changes in mean wind speed are positively correlated to the NAO index, the correlation between (primarily temperature driven) peak demands and the NAO index is negative. This implies that, for example, the sum of demand in GB and Scandinavia net GB wind generation will be more sensitive to the NAO index than either demand or wind output separately.

It may be the case that a significant proportion of demand net wind is to be met by Scandinavian hydro power generators, but during the spring their ability to do so depends on how much melt water they contain, which again is highly dependent on the NAO state, so sensitivity increases further. Since an increased reliance on interconnectors would be quite likely in a future in which radical decarbonisation is achieved, this highlights the importance of incorporating climatological variability into the modelling of renewable resource availabilities.

2.5. Alternative Frameworks for Non-Stationarity

This section will examine alternative frameworks for representing the types of variability discussed in sections 2.3 and 2.4, without explicitly employing discreet states such as the GWL classes or NAOI bands. This begins with evidence in the literature that wind speed time series display a property known as long memory. It is stated by Diebold & Inoueb [85] that long memory is intimately related to regime switching models, including simple mixture models and Markov-switching models, so long as the regimes are similar. Additionally, section 2.5.2 reports on a feature of wind which is very important to capture - conditional heteroskedasticity.

2.5.1. Long Memory

A long memory stochastic process is essentially one which displays persistence (of e.g. higher or lower than average values) on time-scales longer than can be expected by a regular, 'short memory', process. In the time domain, the presence of long memory property is evident from the asymptotic behaviour of a sampe's autocorrelation function, as will be described further in Chapter 3. Auto-correlation functions decay more slowly at large lags for long memory processes and, assuming covariance stationarity, the behaviour is

$$\rho_X(k) \propto k^{2\mathrm{d}-1} \text{ as } k \to \infty, \tag{2.5}$$

where l is the integer time lag and d is the long memory parameter, falling within the range $0 \leq d < 0.5$. The greater the value of d, the greater the process' behaviour differs from a stationary short memory process.

In the frequency domain, long memory manifests as a pole at the origin of the (theoretical) spectral density $f_{\omega,X}(\omega)$, i.e.

$$f_{\omega,X}(\omega) \propto \omega^{-2d} \text{ as } \omega \to 0^{+}. \tag{2.6}$$

For a finite sample, the pole will simply be a peak.

A third definition exists, which clarifies the similarity between regime switching and long memory behaviour. For a random process X_t, the variance of the sum $Y_T = \sum_{t=1}^{T} X_t$ of a long series realisation of length T varies according to

$$var(Y_T) = T^{2d+1}. \tag{2.7}$$

This implies that the variance of sample means varies as T^{2d-1}, so as $d \to 0.5$ the process becomes closer to one in which the variance of sample means is a constant, regardless of sample size. Considered in this way, it is clear that long memory implies self-similarity, i.e.

recurring patterns at every scale, a feature common to e.g. data networks and also meteorological systems.

Self-similarity is also connected to the range of the sum Y_T, i.e. the difference between the maximum and minimum values of Y_T that can be expected for realisations of length T of the process. As described by Beran in [86], if S_T^2 is the sample variance and R_T is the range $max(Y_T) - min(Y_T)$, then for self-similar processes,

$$log\left(E\left[\frac{R}{S}\right]\right) \sim \alpha + H \cdot log(T), \quad \text{with } H = d + 1/2. \tag{2.8}$$

Here α is an unimportant constant and H is known as the self-similarity parameter for the long memory processes.

The fact that H may be greater than 0.5 is known as the Hurst effect, discovered by a hydrologist of that name in 1951. The effect was discovered in the context of calculating the storage capacity required of the Aswan reservoir on the Nile. The goal was to regulate the flow of the Nile, and the sum Y_T was the total flow into the reservoir in time interval T. The empirical discovery was unexpected as it is contrary to results for all stochastic processes usually considered at the time. This is an example of the Joseph effect, in which a high or low value can persist for an extended period before suddenly changing. Such peculiarities in the long-term behaviour of the Nile have been noted since ancient times, the Bible for example reporting on a clustering behaviour with seven years of flooding, followed by seven years of famine due to no flooding (Genesis 41, 29-30).

Such periods are circulation regimes in the language of synoptic types. The long memory framework can be seen as advantageous if it is capable of accurately capturing such behaviour without the need to define somewhat arbitrary classes and clearly defined transitions between them.

A very important study in this context is that of statisticians Haslett and Raftery [88] who were concerned with modelling and assessing Ireland's wind energy potential. They had access to relatively short wind speed time series at locations close to the likely locations of wind generators, as well as longer datasets further away, and consider how to make optimal use of the data. All series were aggregated into daily means to ease computational burdens, without losing too much temporal resolution. The authors examine the behaviour of the simplest estimator for the average wind speeds at locations where only short-run data is available – i.e. the sample mean. They do so for locations where long-run data is actually available, assuming the long-run mean to be the correct value, so that estimator error

distributions could be examined. They also construct a more complex estimator which makes use of long-run data from other locations, the details of which are not relevant here.

It is found that the more complex estimator performs much better than the sample mean, particularly for very short data runs. For n = 20 days, for example, the more former reduces empirical mean squared errors by about 68% compared to the latter. For both cases, theoretical results are obtained for estimator mean squared error as a function of n, assuming only short term persistence, and these are compared with actual empirical errors. The results are striking: theoretical errors are much too small for all values of n, and the relative extent of the difference increases with n. When n = 320, for example, theoretical mean squared errors are too small by a factor of 10 for the sample mean estimator, and a factor of 20 for the other. The long memory property is very clearly present in the wind speed time series.

Periodograms - spectra obtained from finite realisations of a random process, are examined by the authors for further evidence. This is done not for the series but rather AR(9) model errors, with the models fitted following de-trending of the series of their mean annual periodicity. The periodograms for 12 locations across Ireland are very similar, displaying a concentration of power at low frequencies, as expected of a long memory process. Both spectral intensity and frequency are plotted on log scales. At the very lowest frequencies, power is slightly less than expected, which Haslett and Raftery presume is due the finite nature of the samples, rather than a genuine feature of the stochastic process. For each spectrum, with the exception of this effect for the very lowest frequencies, the power drops monotonically over ~3 octaves before becoming characterised by spikes, which group closer together due to the log scale of the plot. Haslett and Raftery do not mention that in each case, one of the first spikes is significantly bigger than all those which follow, in most cases by almost a dB (i.e. factor of 10).

Considering other meteorological time series, Caballero, Jewson and Brix [89] examined three multi-decadal daily time series of mid-latitude near-surface air temperatures, finding long memory dependence in all 3 series, at 95% statistical significance. Evidence that this is the case is presented in the form of more log-log periodograms, showing that the low-frequency region is not flat, as expected for a short memory process, but has a small, constant negative slope, a feature of long memory processes. The authors argue that from a physical point of view, detecting long memory can only be considered interesting if it reveals something about the ‘internal’ workings of the climate system. Thus, they argue, before applying any tests it is necessary to rid the time series from the signature of processes which

are ‘external’ to the climate system. Primary among such externalities is seasonality, where changes in solar forcing give rise to a periodic signal in both mean and variance.

This is in contrast to the view of Bouette et al. [90], who consider the treating of annual seasonality as a deterministic function, to be removed, as a limitation. They take advantage of a generalisation of the definition of a long memory process that rather allows the characteristic seasonalities of the wind resource to be embedded in the model. Processes defined by such models are said to be seasonally persistent. The presence of a pole in the spectrum remains, but it occurs at the seasonal frequency rather than zero.

Seasonally persistent processes are also known as Gagenbauer processes and the spectral peak occurs at the Gagenbauer angular frequency, which must lie within the range $[0,\pi]$. In the time domain, the slow decay of the autocorrelation function for large time lags remains, but it becomes a slowly decreasing envelope for sinusoidal oscillations at the Gagenbauer frequency. A process with both seasonal long memory and short memory behaviour characterised by autoregressive and moving average polynomials is known as a Generealised ARMA, or GARMA, process.

In [88], Haslett and Raftery handle both space dependency and long memory, whereas Bouette et al. attempt to model univariate time series only, without any spatial dependency. The attraction of a model which ties together seasonality and long memory is clear, although Bouette et al. unfortunately do not make a very coherent or compelling argument that for its adoption. They provide a smoothed periodogram for one of the Irish wind datasets, which has not been detrended, in which it is clear that the maximum occurs at the annual frequency. Examining the auto-correlogram for the series, they observe that it decays slowly for moderate lags, a fact that led Haslett and Raftery to fit a long memory model. However, plotting the auto-correlation function for greater lags shows that the slow decay looks sinusoidal; plotting it for lags less than 100 days only shows the beginning of the sine function. To establish this with greater certainty, the authors calculated the spectral density of the auto-correlation function, finding that it exhibits a sharp peak at $2\pi/365$.

The weakness of their approach is that they do not address the nature of the periodogram and autocorrelogram for a process with simple long memory and a deterministic and/or stochastic seasonal trend. Doing so could possibly eliminate this type of process as a means of explaining the empirical periodogram and autocorrelogram. Additionally, they do not extend the empirical autocorrelograms of the deterministically deseasonalised datasets up to

long enough lags to observe if any residual seasonalities are present, which might explain the bigger 1st spike in Haslett and Raftery's periodograms.

The authors also wished to characterise the wind speeds with hourly rather than daily resolution. Data were examined from a weather station at Schipol Airport in the Netherlands, measured between 1951 and 2003. They state that the expected peak at 1/(365 x 24) on the smoothed periodogram of the hourly data can be clearly seen, however there are also peaks of decreasing magnitude for frequencies of 1/24 hours and higher. As a result, modelling the hourly data cannot, they argue, be done with the GARMA process presented above. This would only account for the annual Gagenbauer frequency, and discard the smaller daily frequencies. Their solution is that the GARMA processes can be generalized to take into account multiple Gagenbauer frequencies. The spectrum of a two-factor GARMA process, for example, has poles at two Gagenbauer frequencies. The authors warn that estimating model parameters for such multi-factor GARMA processes is very difficult, so it seems implausible to include frequencies higher than the diurnal.

A major problem with such a model, which they do not consider, is that it does not allow specification of the interaction of the two Gagenbauer frequencies. For example, it cannot capture the fact that when the annual oscillation is at its minimum, i.e. summer, the amplitude of the diurnal oscillation is significantly greater amplitude than at the maximum (winter). It seems that if GARMA is to be the model type chosen for this project, the diurnal seasonality must be removed as a deterministic function of hour of day and day of the year, with annual being the only Gagenbauer frequency. The possibility of additional low frequency factors should not be ruled out, however, if that improves fit to the periodograms.

2.5.2. Heteroskedasticity

As introduced briefly in Chapter 1, a heteroskedastic random process is one in which the variance takes different values during different periods, and it is clear that this is the case for wind speeds. Haslett and Raftery in [88], for example, state that the variance of daily averaged wind speeds in Ireland are correlated to means, but that taking the square roots of the wind speeds stabilises variance.

In addition to seasonal changes, another type of heteroskedasticity present in wind time series, as reported by Tol [91]. At some times, Tol observes, weather prediction is easier than at other times, i.e. for some periods the forecast errors tend to be smaller than the

standard error, while for other periods they tend to be larger. Such periods of large errors can be identified from the time series regardless of model choice, since they correspond more or less to periods of large absolute change in wind speeds from hour to hour.

In this case, the observation of one large (small) prediction error yields the information that the next time-step is more likely than usual to also have a large (small) prediction error. In other words, the variance of errors displays auto-regression and the phenomenon is described as autoregressive conditional heteroskedasticity.

2.6. Chapter Summary

The excellent wind resource in GB is due to the regular passage of cyclonic weather systems across it. However, conditions across GB are in general characterised a combination of weather systems within a larger area. Differential heating of the land and sea leads to a diurnal seasonality, which is stronger in summer.

A gap exists in the spectra of wind speeds for periods of roughly 10 minutes to 1 hour, meaning that variability can be divided into slow synoptic patterns and turbulence. The short-term fluctuations of wind-farms are very small: about 0.1% of plant capacity for 1 second averaged wind speeds, and less than 1% for 1 minute averages. Further, the high-frequency components of power outputs for two separated wind-farms are nearly independent. All this means that hourly averaged wind speeds provide a good representation of the wind resource, since high-frequency fluctuations are smoothed out.

Hourly averaged wind speeds are roughly Weibull distributed – close to the Raleigh distribution when there is no prevailing wind direction, and closer to the Normal distribution when there is. The Weibull distribution along with the Dirac delta function at the origin may be a better description, due to anemometer friction. Better still, in some cases, is a mixed normal distribution.

Winds generally become stronger with increasing height. The relationship between speeds at two heights is often captured by two equations – one involving logarithms and an atmospheric stability term, the other a power law with one coefficient, the exponent α. The latter method is preferred by engineers due to its simplicity. Recent research has shown however that while the power law approach is fine for the calculation of e.g. annual yields, it is not adequate for individual hours. Indeed α values have a broad distribution, including negative values. Important consequences are a reduction of diurnal variations at hub height, and a modest change in Weibull shape parameters with height. In the light of such complexity, this project will not attempt to accurately represent wind shear.

The work of Dr Graham Sinden on the GB wind resource in GB was reported. Sinden produced power outputs from 10m wind speeds, validating his approach by comparing aspects of the derived datasets with those from real, spatially diverse wind fleets in continental Europe. Some of the findings involve temporal patterns and changes in aggregate output, and the results from the synthetic datasets can be compared to these. Other findings relate to the distribution of spatial extents of extremes. A plot of cross-correlation vs. distance, d_p, for a

very large number of location pairs demonstrated that correlations generally follow a $exp(-d_p)$ pattern, although a few % of pairs deviate very significantly from this pattern. The average relationship between demand and wind generation load factor was also thoroughly explored by Sinden, yielding interesting results, although it is admitted that the relationship deviates significantly during individual years.

The process of characterising days according to synoptic types was examined in detail. Such types may correspond to the hidden states in a hierarchical semi-Markov model, although complications would arise, such as the complicated logical filtering that is required for the popular GWL system to satisfy the condition that each type must be at least 3 days in length.

Zachary, Dent and Brayshaw proposed atmospheric circulation-typing, GWL being the best, combined with meteorological data as a means of addressing the data related difficulties described in Chapter 1 - through analysing the weather systems which have historically driven extreme demands. The top two percentiles for hourly demands in England and Wales occurred on only 15 days, of which 10 had 'blocking' GWL types and only one was 'High over Britain'. Concerned that 29 might be a higher than optimal number of types, Brayshaw and Masato reduced them to 7 and 6 types for winter and summer, respectively. If it is decided that a regression or straightforward Markov model is the best choice, rather than a hierarchical one, there is value in exploring options for associating each day in synthetic datasets with one of these types.

Very low frequency changes in the wind resource were then investigated, along with their relationship to the dynamics of large-scale atmospheric circulations, such as the NAO. It was found that annual average wind speeds at a single location tend to vary by a few % year on year, but that windy or calm years may cluster together to form windy periods of longer than a decade. This variability is due mainly to the variability of winters, when large-scale circulation is most influential. The dependence of wind speed on the frequency of westerlies increases gradually from east to west, as intuition would suggest, with more complexity in southern England.

The low frequency variability in the wind resource may be conceived of, alternatively, as long memory. This framework is advantageous if capable of accurately capturing the series' behaviour without the need to define somewhat arbitrary classes and clearly defined transitions between them. The existence of long memory in wind speed series was established

by Haslett and Raftery in 1989, but Bouette et al. argue that wind is better represented as a seasonally persistent, i.e. Gagenbauer process. Some compelling evidence was presented to support the argument, but the authors have not excluded the suitability of processes such as regular long memory combined with stochastic seasonality.

It was noted that the variance of wind speed series display seasonal patterns, which may be corrected for modelling purposes. It has also been noted that observation of a large (small) prediction error yields the information that the next period is more likely to also have a large (small) prediction error. The variance of errors displays autoregression, i.e. is conditional on past values of error, a phenomenon described as autoregressive conditional heteroskedasticity.

A full meteorological context has been presented in this chapter, which has stimulated ideas about the choice of model type and structure. It has also enabled the relationship between wind availability and demand to be understood, to some extent. The next chapter continues by ‘preparing the ground’, mathematically, for a comprehensive survey of wind speed modelling work conducted by others to date in Chapter 4.

Chapter 3

Relevant Time Series Theory

This chapter presents an overview of the core mathematics relating to time series modelling, in order to provide a reference that will facilitate discussion and analysis in subsequent chapters.

It was stated in Chapter 1 that previous wind modelling work at the University of Bath, the result of which was the Bath Wind Model, is the starting point for the model to be developed in this project. The Bath Wind Model is a 20 variable 4^{th} order vector autoregressive model, or VAR(4), combined with fixed wind speed up-scaling. This is an example of the Autoregressive Moving Average (ARMA) model type, also referred to as Box-Jenkins models after statisticians George Box and Gwilym Jenkins, who proposed the model fitting methodology presented here in the 1^{st} edition of [92]. ARMA models, and relevant extensions, are therefore the main focus of this chapter. In addition to basic principles, this chapter discusses the identification of the best model order, spectral properties, the estimation of parameters and model validation. Sources for this Chapter which cannot be associated with only one specific aspect include texts from the following authors: Brockwell and Davis [93]; He, Novac and Glasmeyer [94]; Box, Jenkins and Reinsel [92]; Janacek [95]; and Cryer and Chan [96].

3.1. Basic Principles of ARMA Models

3.1.1. The Concept

Time series analysis is concerned with understanding the mechanisms that give rise to observed sequential data series. It tries to model the series' behaviour and forecast future values based on previous values and, in some cases, external factors. The observed sequence $\{x_1, x_2, \ldots, x_n\}$ is seen as a single realisation of the random process {..., X_{-t}, ..., X_1, X_2, ..., X_t, ...}, which may be expressed simply as $\{X_t\}$ and is defined by a time series model. For Autoregressive Moving Average (ARMA) models, an assumption is made that the value of a series at time t, X_t, depends *linearly* on its previous values (the deterministic part) and on past and present random disturbances (the stochastic part). We may also extend to a multivariate/ vector case, i.e. a collection of concurrent data series, where the components may also be dependent on each other's past values and disturbances.

Returning to the univariate case, if X_t depends on its previous p values and the previous q innovations, we may write:

$$\Phi(B)X_t = \Theta(B)Z_t, \tag{3.1}$$

where $\Phi(x) = 1 - \varphi_1 x - \varphi_2 x^2 - \ldots - \varphi_p x^p$, and $\Theta(x) = 1 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_q x^q$ are known as the autoregressive and moving average polynomials, respectively. Here the $\{\varphi_i\}$ and $\{\theta_i\}$ are real constants; the innovations $\{Z_t\}$ constitute an uncorrelated, zero-mean, usually Gaussian random process with variance σ_z^2; and B is the backward-shift operator defined by the property $B^j X_t = X_{t-j}$. Since the random innovations represent the difference between the actual values of the variables X_t in a particular realisation and the best forecast for them at time $t-1$, they may described as errors, or residuals when fitting a model.

When q is zero, the process is said to be purely autoregressive, and is given by an AR(p) model. Conversely, when p is zero the process is purely moving average, and given by an MA(q) model. When neither are zero, the process is said to be mixed.

In the multivariate case, the AR and MA polynomials are labelled $\mathbf{\Phi}(x)$ and $\mathbf{\Theta}(x)$ respectively. The regression coefficients become matrices, and $\underline{X}_t$ and $\underline{Z}_t$ are vectors. The zero-mean white noise is characterised by a covariance matrix, $\mathbf{\Sigma}_Z$, sometimes assumed to be diagonal, i.e. no noise correlations between sites.

As stated in Chapter 1, ARMA models are essentially linear filters acting on white noise. Further, MA models are finite impulse response filters, since the effect of an impulse Z_t will only be reflected for a finite (q) number of steps. In contrast, AR and ARMA models are

infinite impulse response filters, since 'internal feedback', i.e. the autoregression, mean that an impulse is felt to some extent indefinitely.

In order to make inferences from a single realisation of the random process of interest, ARMA models assume stationarity – i.e. that the statistical properties of the process are time-invariant. More precisely, for ARMA modelling we need only assume weak stationarity – that $\mu(t) = \mu$ and $\sigma(t) = \sigma$, so that the covariance $\gamma(X_{t-s}, X_{t-r})$, defined as $E((X_{t-s} - \mu)(X_{t-r} - \mu))$ is a function of $(s - r)$ only, known as lag, k. This allows definition of consistent covariance and correlation functions, given respectively by

$$\gamma_X(k) = E((X_{t+k} - \mu)(X_t - \mu)), \tag{3.2}$$

$$\rho_X(k) = \gamma_X(k)/\sigma^2 = \gamma_X(k)/\gamma_X(0). \tag{3.3}$$

The multivariate covariance function is defined as $\boldsymbol{\Gamma_X}(k) = E\left(\underline{X}_{t+k} - \underline{\mu}\right).(\underline{X}_t - \underline{\mu})^T$, i.e. a column vector followed by a row vector to produce a square matrix. The correlation function is defined similarly, with matrix elements given by

$$\boldsymbol{\rho_X}(k)_{ij} = \boldsymbol{\Gamma_X}(k)_{ij}/\sqrt{\boldsymbol{\Gamma_X}(k)_{ij}\boldsymbol{\Gamma_X}(k)_{ij}}, \text{ or}$$

$$\boldsymbol{\rho_X}(k) = \boldsymbol{V^{-1/2}}.\boldsymbol{\Gamma_X}(k).\boldsymbol{V^{-1/2}}, \quad \text{where } \boldsymbol{V} = diag(\boldsymbol{\Gamma}(0)_{11}, \ldots, \boldsymbol{\Gamma}(0)_{kk}). \tag{3.4}$$

Note that while the univariate correlation function is symmetric, $\boldsymbol{\Gamma}(k) = \boldsymbol{\Gamma}(-k)^T$.

From now on the presentation shall be simplified by assuming all series have a mean value of zero. This does not involve any loss of generality, since a non-zero mean series can always be transformed by simple subtraction of the mean and adding it back when modelling is over.

An univariate ARMA model is stationary if and only if all the zeros of $\Phi(x)$ lie outside the unit circle in the complex plane. For multivariate models to be stationary, the roots of $det(\boldsymbol{\Phi}(x)) = 0$ must lie outside unit circle. In order to caclulate the latter, the entire $t - 1$ matrix is multiplied with x, the $t - 2$ matrix with x^2 etc – where x is a scalar. Each scalar that makes the determinant of the sum of these matrices zero must be outside the unit circle. Stationary ARMA models can be expressed as MA(∞) models – by inverting the AR polynomial and expanding as:

$$X_t = \Phi(B)^{-1}\Theta(B)Z_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}. \tag{3.5}$$

Similarly, an ARMA model may be expressed as an AR(∞) model, as long as all the zeros of the MA polynomial lie outside the unit circle, or in the multivariate case if the the roots of $det(\boldsymbol{\Theta}(x)) = 0$ lie outside the unit circle. A model satisfying this condition is said to

be invertible. It is possible, in practice, to convert suitable ARMA models into finite AR or MA models by truncating the infinite series when the coefficients, or coefficient matrix determinants, have become suitably small. This is often the case with only 5 or 6 terms.

The auto-correlation function (ACF) for an ARMA process is it's time-lagged auto-correlation as a function of lag, k, and is determined by its polynomial coefficients. For univariate AR(p) processes, the theoretical value of the ACF can be calculated through solving a set of simultaneous equations, known as the Yule-Walker equations. These are obtained by multiplying both sides of (3.1) by X_{t-k}, for k = 1, 2, ..., n -1, rearranging and taking expectations. For the AR(1) model, for example, this yields $\gamma(k) = \varphi_1{}^k \gamma(0)$. They are presented formally in section 3.3.2, equation 3.43.

The ACF of all AR models behave similarly: they decay with increasing lag, but do not fall abruptly to zero at any point. When the AR polynomial has complex roots the correlation function oscillates between positive and negative values, within a decaying envelope. A process with a root close to the unit circle will decay slowly. For MA models, the covariance decays in a more linear manner and is zero for all lags greater than the model order, q.

The partial correlation of X_t and X_{t-k} is the correlation between them when all series values in the intervening time are assumed to be fixed. For example, the partial correlation with a lag of 3 is $E(X_t \, . X_{t-3} \,|\, X_{t-1},\ X_{t-2})$, normalised by $\gamma(0)$. For a known AR(p) model these can be obtained using the Levinson-Durbin algorithm, an iterative method for solving the Yule-Walker equations. For all AR models, partial correlations are smaller than regular correlations for $k \leq p$ and fall abruptly to zero for $k > p$. This is not the case for MA models, since fixing intermediate values actually has the effect of making correlations decay smoothly to infinity, with no cut off at $k > p$. A clear duality exists between purely AR and MA models with regards to regular and partial correlations. As a result, examination of plots of these functions for finite samples, known as auto-correlograms and partial-auto-correlograms, reveals a great deal about the model from which they are derived. Doing so using correlation estimators drawn from data is central to the model identification process.

3.1.2. ARMA Models in the Spectral Domain

While the covariance function is the defining characteristic of a process in the time domain, in the frequency domain analysis is based upon the spectral density function. The relationship between it and the covariance function is that they are Fourier transforms of each

other, the (continuous) power spectral density function for a univariate process $\{X_t\}$ may be defined as:

$$f_{X,\omega}(\omega) = 1/2\pi \sum_{k=-\infty}^{\infty} \gamma_X(k) e^{-ik\omega}. \quad (3.6)$$

There is an assumption here that $\gamma_X(k)$ is absolutely summable (which is not the case for long memory processes at zero or Gagenbauer frequency). Since the correlation function is symmetric, all the properties of the spectrum may be found by examining the function over the range $0 \leq \omega \leq \pi$, and we can re-write the above as

$$f_{X,\omega}(\omega) = 1/2\pi \left[\gamma_X(0) + \sum_{k=1}^{\infty} \gamma_X(k)\cos(k\omega)\right]. \quad (3.7)$$

For a multivariate series the spectrum is simply defined by replacing the covariances with covariance matrices, to give a spectral density matrix. The i^{th} diagonal element is known as the auto-spectrum of the series i, while the off diagonal elements give the cross spectra. One significant difference from the univariate case is that since $\boldsymbol{\Gamma_X}(k)$ is not identical to $\boldsymbol{\Gamma_X}(-k)$, cross spectra are in general complex. As a result, several new spectrum components can now be defined, but are not relevant here.

For a process $\{X_t\}$ derived from filtering another process $\{Y_t\}$ according to

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}, \quad (3.8)$$

it can be derived from (3.6) that its spectral density function $f_{\omega,X}(\omega)$ is related to the function for $\{Y_t\}$, $f_{\omega,Y}(\omega)$, through a power transfer function $T(\omega)$ according to

$$f_{X,\omega}(\omega) = T(\omega) f_{Y,\omega}(\omega) = \left|\psi\left(e^{-i\omega}\right)\right| f_{Y,\omega}(\omega), \text{ where } \psi\left(e^{-i\omega}\right) = \sum_{j=-\infty}^{\infty} \psi_j e^{-i\omega j}. \quad (3.9)$$

When $\{Y_t\} = \{Z_t\}$, i.e. a white noise process with variance σ_z^2, $f_{Y,\omega}(\omega) = \sigma_z^2/2\pi$, since the contribution from each frequency in the range (-π, π) to the total variance is identical.

Since a process defined by a MA(q) model is the result of passing a white noise process through a filter of weights [1, θ_1, θ_2, ... , θ_q], the transfer function in this case is $\left|\theta\left(e^{-i\omega}\right)\right|^2$ and the spectrum for the MA process is simply the transfer function for the filter, scaled by the white noise spectrum $\sigma_z^2/2\pi$. Extending to mixed ARMA processes, spectra are similarly given by

$$f_{X,\omega}(\omega) = \sigma_z^2 \left|\theta\left(e^{-i\omega}\right)\right|^2 / 2\pi \left|\varphi\left(e^{-i\omega}\right)\right|^2. \quad (3.10)$$

For the simple example case of an AR(1) model, the spectrum is therefore

$$f_{X,\omega}(\omega) = \sigma_z^2/2\pi(1 - 2\varphi_1\cos(\omega) + \varphi_1^2). \quad (3.11)$$

For multivariate models, with AR and MA matrix polynomials $\boldsymbol{\Phi}(x)$ and $\boldsymbol{\Theta}(x)$ respectively, the spectral density matrix is given by

$$\boldsymbol{f}_{X,\omega}(\omega) = \left({}^{1}/_{2\pi}\right)\boldsymbol{\Phi^{-1}}\left(e^{-i\omega}\right)\boldsymbol{\Theta}\left(e^{-i\omega}\right)\boldsymbol{\Sigma}_{z}\,\boldsymbol{\Theta}'\left(e^{i\omega}\right)\boldsymbol{\Phi}'^{-1}\left(e^{i\omega}\right). \quad (3.12)$$

In practice, it is often only necessary to calculate the spectral density on a set of discrete Fourier frequencies $\omega_j = 2\pi j/n$ for j = 0, 1, …, n - 1. In this research, datasets that are decades in length imply n will be of the order of hundreds of thousands, leading to a very high resolution. In order to calculate the discrete spectrum, it is not necessary to first obtain the correlation function or model coefficients, rather it may be calculated directly from the data using

$$\hat{f}(\omega) = (1/2\pi n)\left|x(\omega_j)\right|^2, \text{ where } x(\omega_j) = \sum_{t=1}^{n} x_t e^{i\omega_j}. \quad (3.13)$$

Spectral density plots are invariably noisy, with $\hat{f}(w)$ not a consistent estimator – i.e. the variance does not vanish even as $n \to \infty$. For this reason, effective smoothing of the spectrum is crucial.

3.1.3. Seasonal Processes

Stochastic processes often display periodic variations, and in many cases it is possible to capture these with deterministic functions, usually a superposition of sines and cosines, and remove them before model fitting. For some processes however the seasonal variations are stochastic, and deterministic removal is not possible. For such processes we generally require extension to seasonal ARMA (SARMA) models. Such models are multiplicative seasonal models with polynomials for both B and the seasonal differencing operator B^s, and a SARMA$(p,q)(P,Q)_s$ model may be written as:

$$\Phi_s(B^s)\Phi(B)X_t = \Theta(B^s)\theta(B)Z_t. \quad (3.14)$$

Extension to multivariate models is conceptually trivial, with the seasonal polynomials replaced with matrix polynomials.

It follows that correlations at lags which are multiples of the seasonal period are local extrema in the correlograms, although not necessarily sharp peaks. For a process such as wind with an annual seasonality the peak will be spread out, with the correlation for a lag of 8760 hours, for example, probably very similar to that for nearby 8766 hours.

It is also worth considering the spectra of SARMA models, with the model $(1 - \Phi B^6)(1 - \varphi B)X_t = Z_t$, for a process with a seasonal period of 6 hours, chosen here as an example.

Applying (3.10) to this model gives the spectral density

$$f_{X,\omega}(\omega) = \sigma_z^2(1 - 2\varphi cos[\omega] + \varphi^2)(1 - 2\Phi\cos[6\omega] + \Phi^2)/2\pi. \quad (3.15)$$

Within the range of interest for ω, 0 to π, the term $(1 - 2\Phi\cos[6\omega] + \Phi^2)$ means that the spectrum has 3 local maxima and minima, contained within an envelope increasing from $(1-\varphi)^2$ to $1+\varphi^2$. However if wind speed is a SARMA process, with a seasonal period of 8766 hours (as well as 24 hours), there would be a total of 4383 maxima and minima within the range of interest. So close together, these appear as spikes, and a plot of $f_{X,\omega}(\omega)$ would look extremely 'messy'. Indeed, we do notice in Chapters 6, 7 and 8 that periodograms (spectral density plots of the observed series) are indeed very 'spiky'.

3.1.4. Differencing

The requirement of stationarity for ARMA models is rather restrictive, in that processes we wish to model often display trends, e.g. a gradual increase over time. Sometimes the non-constant mean $\mu(t)$ may be described by a deterministic function, such as a polynomial in t, along with a possible seasonal element, that can be removed easily to get a zero-mean process. This is not always the case, as some processes seem to have no fixed mean or regular patterns for it. Such processes may be homogenous, in the sense that any segment seems to behave very much like any other, with the only exception that their means are different. These are in fact models in which not all roots of the AR polynomial lie outside the unit circle, rather d of them are exactly 1. For multivariate models of this kind there are unity roots for $det(\mathbf{\Phi}(x))$.

Such models have the defining equation:

$$\Phi(B)(1-B)^d X_t = \Theta(B)Z_t\,. \tag{3.16}$$

The operation $(1-B)$ is known as differencing, and performing it d times transforms the series into a stationary one. An ARMA model which includes the differencing operator $(1-B)$ is known as an Autoregressive Integrated Moving Average model, ARIMA(p,d,q). Looking back at (3.11) it can be seen that if φ_1 ~ 1, then as $\omega \to 0$, $f_{X,\omega}(\omega) \to \infty$. That is, for processes close to requiring differencing, one may get very large spectral intensities at the low frequency range. This implies that short memory processes with a root close to unity and long memory processes are closely related and their spectra can be very similar for finite samples.

Extending to the multivariate case, an ARIMA processes may be expressed as

$$\mathbf{\Phi}(B)\mathbf{\nabla^d}\underline{X}_t = \mathbf{\Theta}(B)\underline{Z}_t \tag{3.17}$$

where $\mathbf{\nabla^d}$ is a diagonal matrix with elements $(1-B)^d$.

Processes also exist where D_s roots of a seasonal autoregressive polynomial are exactly 1, indicating a requirement for differencing with the seasonal operator $(1 - B^s)^{D_s}$. This leads to the definition of a very general model: SARIMA$(P, D_s, Q)(p, d, q)$, defined by

$$\Phi(B^s)\varphi(B)(1 - B^s)^{D_s}(1 - B)^d X_t = \Theta(B^s)\theta(B)Z_t. \quad (3.18)$$

Extension to multivariate models is conceptually straightforward, with both the regular and seasonal differencing operators being replaced by diagonal matrices.

3.1.5. Transformation of Data

For a variety of reasons, including model estimation and ease of simulation, it is desirable to be able to assume that the process of interest is normally distributed. The simplest way of doing so, when beginning with a process such as wind that is clearly not Gaussian, is to apply the Box-Cox transformation to the data, defined as

$$y(\lambda_{BC}) = (y^{\lambda_{BC}} - 1)/\lambda_{BC}, \text{ for } \lambda_{BC} \neq 0, \qquad y(\lambda_{BC}) = ln(y_{BC}), \text{ for } \lambda_{BC} = 0. \quad (3.19)$$

Here it is assumed that some value of the Box-Cox coefficient λ_{BC} can be found which renders the probability distribution acceptably similar to the normal, and for multivariate data this value may have to be a compromise. Techniques used in the literature to establish the optimal choice of λ for wind speed series are discussed in Chapter 4.

A more precise alternative method is presented by Monbet, Ailliot and Prevosto [97], known as the normal score transformation. This begins with a random variable Y of any distribution, and transforms it into the Gaussian variable X using

$$x = N^{-1}(F_Y(y)) \quad (3.20)$$

where F_Y is the CDF of Y and N is the standard normal CDF. For a multivariate process with variables $\underline{Y} = (Y_1, \cdots, Y_n)$, one may independently apply the transformation on the various components, such that

$$(x_1, \cdots, x_n) = (N^{-1}F_{Y_1}(y_1), \cdots, N^{-1}F_{Y_1}(y_1)) \quad (3.21)$$

Where F_{Y_i} is the CDF of $\{Y_i\}$. A limitation of this process, as reported by the authors, is that when a strong dependence exists between the components of the process, this transformation does not restore the joint distribution. The authors suggest use of the Rozenblatt transformation:

$$(x_1, \cdots, x_n) = (N^{-1}F_{Y_1}(y_1), N^{-1}F_{Y_2|Y_1=y_1}(y_2), \dots, N^{-1}F_{Y_n|Y_1=y_1,\dots,Y_{n-1}=y_{n-1}}(y_n)). \quad (3.22)$$

This seems rather impractical for 20 variables, where e.g. the calculation of the conditional CDF for zone 20, conditional on specific values for all other 19 zones, would be based on an extremely small percentage of the dataset.

3.2. ARMA Model Extensions

3.2.1. Long Memory

In Chapter 2 it was shown that the wind speed resource displays long memory, with arguments in favour of seasonal long memory. Long memory can be modelled through differenced ARMA models, but with the integer differencing order d replaced with the long memory parameter introduced in Chapter 2, with the range $0 \leq d < 0.5$. Processes with an ARMA structure but requiring differencing by the non-integer parameter in order to be rendered stationary are known as fractionally integrated, or ARFIMA processes. They were first proposed in 1980 by Granger & Joyeux [98] and Hosking [99]. The concept of non-integer differencing is rather counterintuitive, but makes more sense if the operator is expanded as an infinite series in B, using Newton's generalised binomial theorem, with truncation at some suitably large number. The theorem states that

$$(x + y)^{r_b} = \sum_{k=0}^{\infty} \binom{r_b}{k} x^{r_b - k} y^k, \text{ where } \binom{r_b}{k} = (r_b - 1)\ldots(r_b - k + 1)/k!, \tag{3.23}$$

and raising $x = 1$ to non-integer powers is not a problem. If we take, for example, d = 0.1 this gives

$$(1 - B)^{0.1} X_t = (1 - 0.1B - 0.045\, B^2 - 0.0285B^3 + \boldsymbol{O}(B^4))X_t. \tag{3.24}$$

For simulation of such a process, the operator may be moved over to the right hand side of the model definition to give

$$\Phi(B)X_t = \Theta(B)\left(1 - B\right)^{-\delta} Z_t. \tag{3.25}$$

Looking again at the example value of d = 0.1, the operator may be expanded as

$$(1 - B)^{-0.1} = 1 + 0.1B + 0.055B^2 + 0.0299B^3 + \boldsymbol{O}(B^4). \tag{3.26}$$

The re-arrangement of equation 3.19 is not possible in the multivariate case due to the non-commutativity of matrix operators, so we are left with

$$\underline{X}_t = \boldsymbol{\nabla}^{-d}\boldsymbol{\Phi}(B)^{-1}\,\boldsymbol{\Theta}(B)\underline{Z}_t\,. \tag{3.27}$$

For the more general case of seasonal long memory it is necessary to introduce a new parameter to represent the seasonal frequency, ω_G. In fact the differencing operator for a Gagenbauer process is

$$\nabla^{\delta,v} = (1 - 2vB + B^2)^{\delta}, \tag{3.28}$$

where $v = cos(\omega_G)$, such that the defining equation for a GARMA(p, v, δ, q) model, proposed in 1989 by Gray, Zhang and Woodward [100] is

$$\Phi(B)(1 - 2vB + B^2)^{\delta} X_t = \Theta(B)Z_t. \tag{3.29}$$

When $\omega_G = 0$, the differencing operator reduces to $(1 - B)^{2\delta}$ and the process is regular ARFIMA. For a GARMA process defined by equation 3.29 to be stationary and invertible we must have either $|v| = 1$ and $-¼ < \delta < ¼$ or $|v| < 1$ and $-½ < \delta < ½$.

With this new operator, the binomial expansion theory can no longer be used to apply differencing. However an alternative expansion exists, comprising of Gagenbauer polynomials $C_j(\delta, v)$, i.e.

$$(1 - 2vx + x^2)^{-\delta} = \sum_{j \geq 0} C_j(\delta, v)\, x^j. \tag{3.30}$$

The Gagenbauer polynomials $C_k(\delta, v)$ may be defined either through a long analytical expression or, more usefully, through the recursive algorithm:

$$C_0(\delta, v) = 1, \qquad C_1(\delta, v) = 2\delta v \tag{3.31}$$

$$\forall\, j > 1,\ C_j(\delta, v) = 2v\left(\frac{\delta - 1}{j} + 1\right) C_{j-1}(\delta, v) - \left(\frac{2(\delta - 1)}{j} + 1\right) C_{j-2}(\delta, v). \tag{3.32}$$

Figure 3.1 shows the 1st 20 polynomials for the annual frequency and δ = 0.1, expanded on both the left and right hand sides of 3.29, while figure 3.2 shows the behaviour up to very large index values. The latter figure reveals the annual periodicity present for both expansions but that the l.h.s. expansion, barely visible, decays much more quickly.

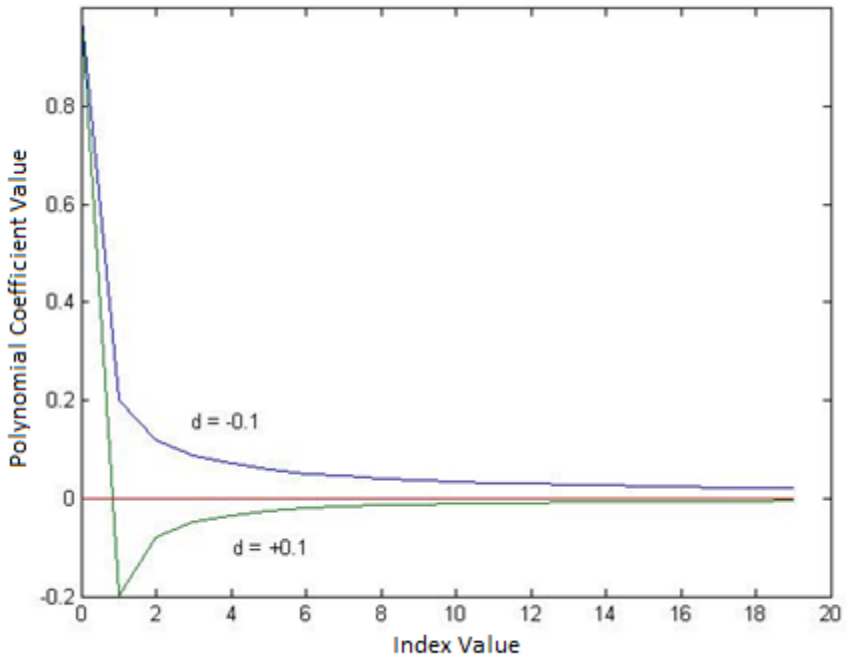

Figure 3.1. The 1st 20 Gagenbauer Polynomials, |d| = 0.1, annual frequency.

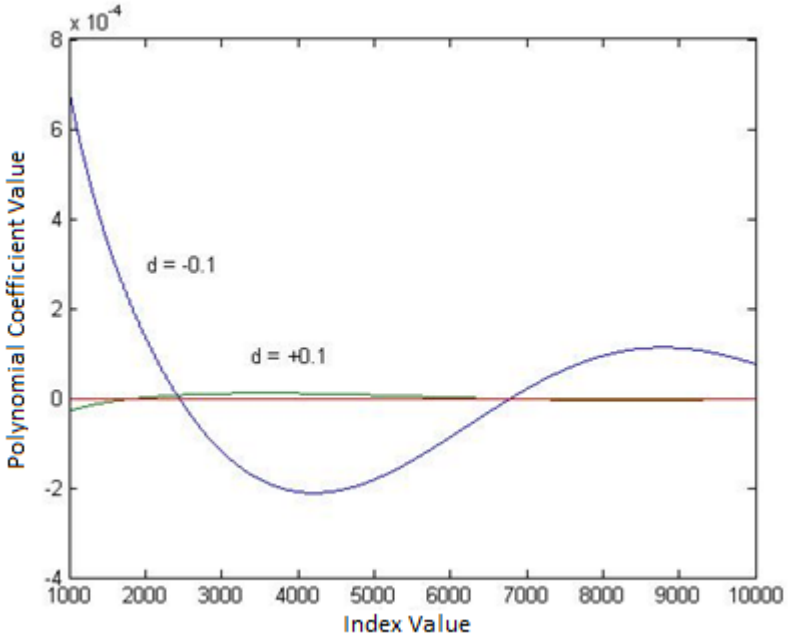

Figure 3.2. The Gagenbauer Polynomials, |d| = 0.1, annual frequency, large index values.

Estimating v is usually straightforward as it is simply the cosine of the frequency for which the periodogram is a maximum, which should be 2π/8766 for hourly averaged wind speeds. The parameter δ dictates how steeply the spectral intensity function rises close to the unbounded frequency, and as a result one can construct an initial estimator for it based upon the periodogram in this region. Extension to multivariate models is conceptually easy, with the differencing operator replaced by a diagonal matrix as in equation 3.17. Both differencing parameters are specific to each zone, although clearly in some cases we assume that all zones share a common value for v.

An alternative means of combining seasonality and fractional differencing is with a model type proposed by Gil-Alana [101]:

$$(1 - B^s)^d Y_t = X_t, \tag{3.33}$$

where d is again the long memory parameter and X_t is a stationary (seasonal or non-seasonal) ARMA process. This is described as a seasonal long memory process, because of the strong association between observations at seasonal lags – widely separated in the case of wind. If $d = 0$, $Y_t = X_t$, a stationary model, while if $d = 1$ we have a seasonal unit root model. This does not appear to be as good a choice for wind speeds as GARMA, since the annual period is so long compared to the hourly time-step.

3.2.2. Conditional Heteroskedasticity

The data transformation described above will surely have the effect of stabilising variance, since higher values will be brought closer together, but as discussed in Chapter 2 this might not entirely eliminate the conditional nature of variance. A suitable extension to an ARMA/ SARIMA/ GARMA model may be necessary to reproduce the behaviour of wind. Such conditional heteroskedasticity models were first developed by econometricists during the 1980s [102], as financial time series tend to be heteroskedastic. The development of such models in fact lead to the 2003 Nobel Prize for Economics being awarded to two of the main contributors, Engle and Granger. For such models, the innovations process $\{Z_t\}$ is given by

$$Z_t = h_t^{1/2}\xi_t \tag{3.34}$$

where ξ_t is an i.i.d. white noise process with zero mean and unit variance, the simplest being $N(0,1)$. Clearly $h_t = \sigma^2(t)$, the conditional variance for the innovation series. A multitude of model types have now been developed for h_t, as described in [103]. A selection of such models, all of which could potentially be suitable for capturing the variance dynamics of wind, are described below. Some symbols have been changed from those appearing in the original model definitions in order to minimise confusion, given the large number of equations presented in this chapter.

The simplest are Autoregressive Conditional Heteroskedasticity (ARCH) models, in which the conditional variance depends on the last few innovation values. More useful is the generalised ARCH type, GARCH, in which h_t depends on its own past values (as an infinite response filter), as well as past innovation values (as a finite response filter), such that the GARCH(r,s) model has h_t defined by

$$h_t = \alpha_0 + \sum_{i=1}^{r}\alpha_i z_{t-i}^2 + \sum_{j=1}^{s}\beta_j h_t. \tag{3.35}$$

Low values of r and s are adequate for most applications, with GARCH(1,1) often chosen. Among the variety of extensions to the basic GARCH model, those of interest to this research involve the introduction of asymmetry, i.e. allowing negative changes in the series to have a stronger or weaker effect on variance than positive changes. One example of such a model is the bilinear GARCH, BL-GARCH(p,q,r), introduced by Storti & Vitale [104]. In this case, the conditional variance is given by

$$h_t = \alpha_0 + \sum_{i=1}^{r}\alpha_i z_{t-i}^2 + \sum_{j=1}^{s}\beta_j h_{t-j} + \sum_{k=1}^{w}\gamma_k z_{t-k} h_{t-k}^{1/2}, \tag{3.36}$$

i.e. there are additional terms introducing a very flexible asymmetry into the influence of positive and negative innovations.

Another model example involving asymmetry is the Threshold GARCH, TGARCH, in which the α_i coefficient values change with the sign of the innovation. Introduced by Zakoian [105], this model is limited to being 1st order in z_t and $h_t^{1/2}$ and has the form

$$h_t^{1/2} = \alpha_0 + \alpha^+ z_{t-1}^+ + \alpha^- z_{t-1}^- + \beta h_{t-1}^{1/2}, \tag{3.37}$$

where $z_t^+ = z_t$ for $z_t > 0, = 0$ otherwise and $z_t^- = z_t$ for $z_t \leq 0, = 0$ otherwise.

A very general model was introduced by Ding, Granger & Engle [106] – the asymmetric power ARCH, APARCH(r, s, λ, ι) model:

$$(h_t^{1/2})^\iota = \alpha_0 + \sum_{i=1}^{r} \alpha_i (|z_{t-i}| - \lambda z_{t-i})^\iota + \sum_{j=1}^{s} \beta_j (h_{t-j}^{1/2})^\iota. \tag{3.38}$$

The term power in the model name clearly refers to the fact that the conditional variance may be raised to any real positive power ι in the defining equation. The parameter λ dictates the direction and extent of asymmetry and has the range -1 < λ < 1. As it is so general, the APARCH model contains several ARCH-based models as special cases, some of which are not described here.

In general, the dynamics of innovation variance are not connected to that of the series mean, and the parameters that characterise mean behaviour are separate from those which characterise variance. The exception is ARCH-in-mean models in which the model for the conditional mean $\mu_{t|I}$ is extended to include a term involving h_t. For example, an ARMA(p, q) model might be extended to become

$$\Phi(B)X_t = \Theta(B)z_t + \chi_{CM} h_t^{1/2}, \tag{3.39}$$

for some coeffient χ_{CM}, and h_t may have any specification, such as APARCH. It is possible that such an extension is required for wind speed modelling, since there is a strong positive correlation between mean and variance. However it is hoped, for the sake of simplicity, that deterministic seasonal detrending alone can account for this relationship to a satisfactory approximation.

It is desirable that the model can capture the relationship between volatilities at different locations. Interesting dynamic aspects which the model should *ideally* capture include:

- Any tendency for a volatility change at one location to precede a similar change at another;
- The extent of the effect of a large innovation, or ‘shock’ in econometrics terms, at one location on the volatility at others;
- Any asymmetry in such relations with regard to the direction of the shock; and

- Whether volatility cross-correlations change over time – perhaps increasing during times of generally high volatility.

Such questions about the specification of the dynamics of covariances can be studied directly by using a multivariate model of conditional variances. For an m-variate conditional variance process with innovations $\underline{z}_t$ we define a m x m conditional covariance matrix $\boldsymbol{H}_t$ such that

$$\underline{z}_t = \boldsymbol{H}_t^{1/2} \underline{\xi}_t, \tag{3.40}$$

where is a m x 1 vector of independent white noise processes, with zero mean and unit variances. Since $\boldsymbol{H}_t^{1/2}$ may be any m x m positive definite matrix such that $\boldsymbol{H}_t$ is the correct conditional covariance matrix, it can be obtained by Cholesky factorization – a calculation which might have to take place for every simulated hour, since equations deal with $\boldsymbol{H}_t$.

A great variety of model types exist for $\boldsymbol{H}_t$, categorised by Bauwens, Laurent and Rambouts in [107] into 3 types: (i) direct generalizations of the univariate GARCH model; (ii) linear combinations of univariate GARCH models; and (iii) nonlinear combinations of univariate GARCH models.

A very general multivariate GARCH model, in category (i) above, is VEC(1,1) (the meaning of the abbreviation is not given). In this model, each element of $\boldsymbol{H}_t$ is a linear function of the lagged squared errors and cross-products of errors and lagged values of the elements of $\boldsymbol{H}_t$. The model specification requires the introduction of vectors $\underline{h}_t$ and $\underline{\eta}_t$ which are the lower triangular portion of matrices $\boldsymbol{H}_t$ and $\underline{z}_t \, \underline{z}_t'$, respectively, stacked into $m(m+1)/2$ x 1 vectors. The structure is:

$$\underline{h}_t = \underline{c} + \boldsymbol{A}\underline{\eta}_{t-1} + \boldsymbol{G}\underline{h}_{t-1}, \tag{3.41}$$

where $\boldsymbol{A}$ and $\boldsymbol{G}$ are square parameter matrices and $\underline{c}$ is a parameter vector. The total number of parameters is $m(m+1)(m(m+1)+1)/2$ which for a 20 zone model is an implausible 88,410. One possible simplification, which retains a high level of generality, is to assume the matrices $\boldsymbol{A}$ and $\boldsymbol{G}$ to be diagonal such that the elements of $\underline{h}_t$ depend on only their own lagged values, as well as the lagged values of all covariances. This model requires a somewhat more manageable $m(m+5)/2 = 250$ parameters.

3.3. ARMA Model Fitting

3.3.1. Introduction

This section provides a conceptual introduction to the fitting of ARMA models. A more detailed examination of methodologies for fitting models with specific extensions such as seasonal long memory and asymmetric conditional heteroskedasticity is reserved for Chapter 7, following an exploration of which model type is best suited to replicate the wind speed field. For a given model structure/ type, the process can be divided into three stages: preliminary identification of the best model order(s), parameter estimation for that choice and testing whether the order choice is optimal. These activities are closely related and interdependent, for example tools used in model idetification rely on an parameter estimators for the model investigated. There's no guarranteed method for getting the right model – the process is iterative and gaining experience and developing judgement are necessary.

The relationship between sample estimators and the property they are estimating is usually very straightforward. For example, given a datset of length n, we define the sample covariance estimator at lag k as

$$\widehat{Y_X}(k) = \sum_{t=1}^{n-k}(X_{t+k} - \bar{X})(X_t - \bar{X})/n \tag{3.42}$$

where $\bar{X}$ is the sample mean. Note that the denomionator is always n, despite there being fewer than n terms in the summation for lags greater than zero, which may have a significant impact for larger lags where $k \sim n$. Sample correlograms are constructed in the obvious way using $\widehat{\rho_X}(k) = \hat{Y}_X(k)/\widehat{Y_X}(0)$, and the main tool for drawing initial conclusions about the model type and order.

A very useful method for order selection is the drawing of boundaries on the sample correlogram whose values are $\pm\, 2/\sqrt{n}$. This is based on the fact that for white noise processes the theoretical correlation function is 0 for all nonzero lags, and the variance of the correlation estimator as $n \to \infty$ is $1/n$. For Gaussian noise, this implies that 95% of the time the white noise correlation estimate for any nonzero lag should fall within the boundaries $\pm\, 1.96/\sqrt{n} \sim \pm\, 2/\sqrt{n}$. As a result, if a sample correlation $\hat{\rho}(i)$ for any process lies outside of this boundary, it is unlikely to be noise, i.e. a genuine correlation probably exists at this lag. This rule can be applied to both auto-correlograms and cross-correlograms, and will be taken here as a good indicator of significance even if the noise is not exactly Gaussian. Sample partial-correlograms are also vital tools, with a definite cutoff indicating a pure AR process, and the $2/\sqrt{n}$ boundaries helping to estabish exactly at which lag such a cutoff takes place. To plot

sample partial correlograms, one must use the Levinson-Durbin algorithm, described below, using sample covariances as an approximation.

3.3.2. The Yule-Walker Equations and Levinson-Durbin Algorithm

It section 3.1.1 the Yule-Walker equations were mentioned as a means of calculating the expectation value of lagged correlations when model parameters are known. They can also be used conversely to estimate those parameters based on sample correlations. In order to fit an AR(p) model, the 1st p simultaneous equations may be expressed in a re-arranged matrix form as

$$\underline{\Phi_p} = \boldsymbol{\Gamma_p^{-1}}\underline{\Gamma_{(p)}} \tag{3.43}$$

where $\underline{\Gamma}_{(\mathrm{p})} = \left[\widehat{\Upsilon_X}(1), \widehat{\Upsilon_X}(2), \ldots, \widehat{\Upsilon_X}(p)\right]^{\mathrm{T}}$, $\underline{\Phi_p} = (\hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_p)^{\mathrm{T}}$ and

$$\boldsymbol{\Gamma_p} = \begin{bmatrix} \widehat{\Upsilon_X}(0) & \cdots & \widehat{\Upsilon_X}(1-p) \\ \vdots & \ddots & \vdots \\ \widehat{\Upsilon_X}(p-1) & \cdots & \widehat{\Upsilon_X}(0) \end{bmatrix}.$$

The Levinson-Durbin algorithm is an iterative method of solving the equations, fitting pure AR models of increasing order to a data set. If we define $\widehat{\varphi}_{m,j}$ to be the j^{th} coefficient of a fitted AR(m) model, and $\widehat{\sigma}_m^2$ the noise variance of that model, the procedure for creating each model is given by the equations:

$$\widehat{\varphi}_{m,m} = \left[\widehat{\Upsilon_X}(m) - \sum_{\mathrm{j}=1}^{\mathrm{m}-1} \widehat{\varphi}_{m-1,j}\, \widehat{\Upsilon}_X(m-j)\right] / \widehat{\sigma}_{m-1}^2, \tag{3.44}$$

$$\widehat{\varphi}_{m,j} = \widehat{\varphi}_{m-1,j} - \widehat{\varphi}_{m,m}\, \widehat{\varphi}_{m-1,m-j}, \quad \text{for } j = 1, 2, \ldots, m\text{-}1 \tag{3.45}$$

$$\widehat{\sigma}_m^2 = \widehat{\sigma}_{m-1}^2\left(1 - \widehat{\varphi}_{m,m}\right) \tag{3.46}$$

The calculation may be started by considering a zeroth order model – i.e. by assuming it is a white noise process. In this case $m = 0$ and the noise variance $\widehat{\sigma}_m^2 = \widehat{\Upsilon}(0)$. Moving on to $m = 1$, the iteration is begun by defining $\widehat{\varphi}_{1,1}$ as $\widehat{\Upsilon_X}(1)/\widehat{\sigma}_0^2$. The second equation is irrelevant in this case, but the third now provides a value for $\widehat{\sigma}_1^2$. Progression to $m = 2$ is now possible, using the 3 equations to obtain values for $\widehat{\varphi}_{2,2}$, $\widehat{\varphi}_{2,1}$ and $\widehat{\sigma}_2^2$. This process may be repeated any number of times. It can be easily shown that $\widehat{\varphi}_{k,k}$ is the sample partial correlation for the lag k and therefore the Levinson-Durbin algorithm must be employed to calculate partial correlations.

This method was generalised to the multivariate case by Whittle [110], who realised that in order to extend the recursion of the univariate case one must fit two auto-regressions simultaneously: a relation expressing $\underline{X_t}$ in terms of its immediate past and another in terms of

its immediate future. Changing the notation slightly, we define matrices $\boldsymbol{A}_{p,k}$ for the VAR(p) model fitted to an m-variate series $\underline{X}_t$ according to

$$\sum_{k=0}^{p} \boldsymbol{A}_{p,k}\, \underline{X}_{t-k} = \underline{Z}_t, \text{ where } \boldsymbol{A}_{\mathrm{p},0} = \boldsymbol{I}_{\boldsymbol{m}}. \tag{3.47}$$

In order to fit VAR models of increasing order to the data, we must define four matrices for each order p:

$$\boldsymbol{V}_p = \sum_{k=0}^{p} \boldsymbol{A}_{\mathrm{p,k}}\, \boldsymbol{\Gamma}_{\boldsymbol{X}}(-k), \quad \boldsymbol{\Delta}_p = \sum_{k=0}^{p} \boldsymbol{A}_{\mathrm{p,k}}\, \boldsymbol{\Gamma}_{\boldsymbol{X}}(p-k+1), \tag{3.48}$$

$$\overline{\boldsymbol{V}}_p = \sum_{k=0}^{p} \overline{\boldsymbol{A}}_{\mathrm{p,k}}\, \boldsymbol{\Gamma}_{\boldsymbol{X}}(k), \quad \overline{\boldsymbol{\Delta}}_p = \sum_{k=0}^{p} \overline{\boldsymbol{A}}_{\mathrm{p,k}}\, \boldsymbol{\Gamma}_{\boldsymbol{X}}(p-k+1). \tag{3.49}$$

Whittle [110] established four recurtion relations between these matrices that allow them to be calculated for an order p model, beginning with the 0th order, and with the true correlation matrices above replaced with sample estmates. The relations are:

$$\boldsymbol{A}_{p+1,p+1} = -\boldsymbol{\Delta}_p\, \overline{\boldsymbol{V}}_p^{-1}, \quad \overline{\boldsymbol{A}}_{p+1,p+1} = -\overline{\boldsymbol{\Delta}}_p\, \boldsymbol{V}_p^{-1} \tag{3.50}$$

$$\boldsymbol{A}_{p+1,k} = \boldsymbol{A}_{p,k} + \boldsymbol{A}_{p+1,p+1}\, \overline{\boldsymbol{A}}_{p,p-k+1}, \quad \overline{\boldsymbol{A}}_{p+1,k} = \overline{\boldsymbol{A}}_{p,k} + \overline{\boldsymbol{A}}_{p+1,p+1}\, \boldsymbol{A}_{p,p-k+1} \quad (j = 1,2,\ldots,p). \tag{3.51}$$

Using the Levinson-Durbin algorithm, one can plot how the noise variance, or the determinant of the noise covariance matrix in the multivariate case, decreases with increasing p. The point at which this curve flattens out gives a good first indication of model order.

3.3.3. Information Criteria and the Hannan-Rissanen Procedure

Whereas the correlograms and plots of reduction in noise are essential in gaining a first impression of the correct order, a more rigorous method is also necessary. This is provided by two information criteria: Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). These involve minimising a penalty function comprising of two terms – one which gets larger as the model order increases and another related to the estimated noise variance, such that minimising the function equates to balancing the risk of under- and over-fitting a model to the data. The criteria are defined as follows:

$$AIC = ln\, \hat{\sigma}_{\mathrm{z}}^2 \; + \; 2(p+q)/n \tag{3.52}$$

$$BIC = \; \ln \hat{\sigma}_{\mathrm{z}}^2 \; + \; (p+q) ln(n)/n. \tag{3.53}$$

The noise variance $\hat{\sigma}_{\mathrm{z}}^2$ must be obtained using a parameter estimation method – such as the methods described above, or by ordinary least squares (OLS) estimation. The AIC and BIC for an m-variate (vector) process are

$$AIC = \ln(\det(\hat{\Sigma}_{\mathrm{z}})) \; + \; 2m^2(p+q)/n \tag{3.54}$$

$$BIC = \ln(\det(\hat{\Sigma}_{\mathrm{z}})) \; + \; m^2(p+q)\ln n\,/n. \tag{3.55}$$

The information criteria may be integrated into a model selection procedure by setting upper limits on the values which p and q may take, based on the initial analysis of correlograms, then evaluating the criteria for every combination of p and q up to these maximum values. For long datasets the BIC is more parsimonious, i.e. it tends to favour lower order models.

Parameter estimation is computationally much simpler for pure AR models than for MA and mixed ARMA models, as the Yule-Walker equations OLS optimisation only apply for AR models. During model fitting, noise innovations become model residuals, errors to be minimised. For MA and ARMA models one is dealing with unobserved quantities (the innovations), which can only be derived given a full set of parameters. It must also be assumed that both the series and the innovations are zero at times $(-\max(p,q)+1)$ to 0, but this has little effect for very long series such as decades of hourly wind speeds. Inefficiency results from the fact that the 1st set of parameters used to obtain the innovations will not be an optimal fit. Specific algorithms exist for the fitting of MA models, but it is not necessary to describe them here.

One efficient procedure for mixed ARMA models, which will be used during this research, is the Hannan-Rissanen procedure. The first stage of this methodology is to use the Levinson-Durbin algorithm to fit models AR(k) to the data, for $i = 1, 2, \ldots, k_{max}$, and evaluate the AIC in each case, to establish the best fit. Since AIC is the least parsimonious of the criteria, k will almost certainly be larger than the final order p. Using the AR parameters $\{\hat{\varphi}_i\}$ from the best fitting model, one can then generate a set of estimated innovations $\{\hat{z}_t\}$ from the data $\{x_t\}$. The next stage, repeated for each combination of $p \leq min(p_{\max}, k)$ and $q \leq q_{max}$, is to define a new set of innovations $\{z_t\}$ as

$$z_t = x_t - \varphi_1 x_{t\text{-}1} - \cdots - \varphi_p x_{t\text{-}p} - \theta_1 \hat{z}_{t\text{-}1} - \cdots - \theta_q \hat{z}_{t\text{-}q}. \quad (3.56)$$

This allows use of the efficient OLS method to estimate the coefficients $\{\varphi_i\}$ and $\{\theta_i\}$. From this we can calculate the white noise variance and hence the BIC for each combination of p and q, and the combination that minimises it is chosen as the best model order. Extension to the multivariate case is again conceptually straightforward, especially if it is assume that the objective of the least squares optimisation is to minimise the variance of residuals on a per zone basis, i.e. without considering cross-covariances. These are however included in the information criteria calculations.

A positive feature of this methodology is that it provides 'runner up' models, in addition to the best fitting. It is very important in the context of this research that models fully

capture spatiotemporal patterns such the typical progress of a weather front across the country. The need to reflect such phenomena makes higher order multivariate models more attractive, and might motivate the choice of second or even third best models with a slightly higher order, if the 1st choice order is low. It must be remembered, however, that higher models imply a lengthier parameter estimation process. Broersen [111] points out that in multivariate ARMA model fitting, the optimum model order for a particular type of process tends to increase for an increasing number of variables.

3.3.4. Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) of parameters for a given model structure and sequential set of observations means calculating the set of parameter values for which the series of observations are most likely to have occurred. Parameters obtained in this way are called the maximum likelihood estimates of the parameters. The method is often favoured as its estimators have the smallest variance of all methodologies. It is standard practice to use parameter values estimated from other methods as initial values for MLE, and as a result estimates not based on maximizing the likelihood function are often regarded as preliminary.

In the context of likelihood estimation, a time series of n normally distributed observations of a single quantity are considered to form an n-dimensional vector: $\underline{x}_n = (x_1, \dots, x_n)'$. This vector is considered to be one realisation of the random vector $\underline{X}_n = (X_1, \dots, X_n)'$, that has a MVN distribution. Such a distribution is characterised by a vector of expectation values $\underline{\mu}_n$ – assumed here to be all zeros, and a covariance matrix $\boldsymbol{\Gamma}_n = E(\underline{X}_n . \underline{X}_n{}')$. Given this matrix, the likelihood function for the observations, i.e. the probability of the vector $\underline{X}_n$ taking the values $\underline{x}_n$, is:

$$L(\boldsymbol{\Gamma}_n) = (2\pi)^{-n/2} (det(\boldsymbol{\Gamma}_n))^{-1/2} exp(-\frac{1}{2}\underline{X}'_n \boldsymbol{\Gamma}_n{}^{-1} \underline{X}_n). \quad (3.57)$$

For the datasets dealt with in this project, n is of the order of 2 x 10^5, implying that calculation of $det(\boldsymbol{\Gamma}_n)$ and $\boldsymbol{\Gamma}_n{}^{-1}$ would be computationally extremely expensive. Fortunately, it can be shown e.g. [93], that this can be avoided by expressing the likelihood function in terms of the innovations, expressed here as $X_j - \hat{X}_j$, where $\hat{X}_j = \mu_{t|I} = E(X_j | X_1, \dots, X_{j-1})$, along with v_{j-1}, the j^{th} component of the diagonal covariance matrix for the innovations. Assuming that $\hat{X}_1 = 0$, the (conditional) likelihood function becomes

$$L(\Gamma_n) = (2\pi)^{-\frac{n}{2}} (v_0 \dots v_{n-1})^{-1/2} exp(-\frac{1}{2}\sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / v_{j-1}). \quad (3.58)$$

With a slight change of notation, the set of AR coefficients may be represented by φ, and the set of MA coefficients by θ. It can be shown, see e.g. [93], through algebraic manipulation of the likelihood function, taking a logarithm and differentiating partially, that for an ARMA model the maximum likelihood estimators $\hat{\phi}, \hat{\theta}$ and $\hat{\sigma}_z^2$ satisfy the equations

$$\sigma_z^2 = n^{-1}S(\hat{\phi}, \hat{\theta}), \quad \text{where} \quad S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2/r_{j-1}, \tag{3.60}$$

and $\hat{\phi}, \hat{\theta}$ are the values of ϕ, θ that minimise

$$l(\phi, \theta) = \ln\left(n^{-1}S(\phi, \theta)\right) + n^{-1}\sum_{j=1}^{n} ln\, r_{j-1}. \tag{3.61}$$

The $\{r_n\}$ are a convenient construct satisfying $E(X_{n+1} - \hat{X}_{n+1})^2 = \sigma_z^2 r_n$. For an AR model, the r_{j-1} are all unity, and so for large n the maximum likelihood estimates are very close to those obtained by ordinary least squares. Although these results are for Gaussian processes, equation 3.59 still holds as a measure of goodness-of-fit for a set of parameters if the data are not Gaussian.

Brockwell and Davis state in [93] that if $\{\underline{X}_t\}$ is a Gaussian m-variate time series with prediction errors $\{\underline{Z}_t\}$, where $\underline{Z}_j$ has the covariance matrix $\boldsymbol{V}_{j-1}$, the probability density for set of specific errors $(\underline{z}_1, \ldots, \underline{z}_n)$ is

$$f_Z(\underline{z}_1, \ldots, \underline{z}_n) = (2\pi)^{-\frac{nm}{2}}\left(\prod_{j=1}^{n}\det(\boldsymbol{V}_{j-1})\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\underline{z}_j'.\boldsymbol{V}_{j-1}^{-1}.\underline{z}_j\right). \tag{3.62}$$

For an AR(p) process with a set of coefficient matrices denoted by $\boldsymbol{\Phi}$ and white noise covariance matrix $\boldsymbol{\Sigma}_\mathbf{Z}$ the likelihood of observations $(\underline{X}_1, \ldots, \underline{X}_n)$ can be expressed as

$$L(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_z) = (2\pi)^{-\frac{nm}{2}}\left(\prod_{j=1}^{n}\det(\boldsymbol{V}_{j-1})\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\underline{Z}_j'.\boldsymbol{V}_{j-1}^{-1}.\underline{Z}_j\right). \tag{3.63}$$

Maximum likelihood estimation is much more challenging for multivariate series – partially because of the potentially very large number of variables involved, and also because it is not possible to compute the maximum likelihood estimator of $\boldsymbol{\Phi}$ independently of $\boldsymbol{\Sigma}_\mathbf{z}$, as can be done for univariate series. The omission of MA coefficients strongly suggests that the maximum likelihood estimation of mixed and MA models is problematic, and this is surely the case with 20 variables. Clearly such problems become even more severe when extensions such as long memory and conditional heteroskedasticity are introduced. It is therefore unsurprising that methods of simplifying such calculations constitute a substantial proportion of many texts of specific ARMA-extension based modelling.

3.3.5. Testing Model Residuals

In order to assess the quality of a fitted model, several statistical tests exist to establish whether the set of model residuals $\{\hat{z}_t\}$ constitute an independent and identically distributed white noise process. The first step is to plot the correlogram for the residuals, along with the $\pm 2/\sqrt{n}$ boundaries. If almost all correlations fall within these boundaries, this is a good indication that the residuals are indeed white noise. For a more rigorous validation one may use:

- The turning points test – for white noise the number of local extrema has an asymptotic normal distribution with mean $2(n-2)/3$ and variance $(16n-29)/90$.
- The sign-difference test – for a random process with no trend, the number of times ($z_t - z_{t-1}$) > 0 has an expectation value of $(n-1)/2$ and variance of $(n+1)/12$.
- The portmanteau statistic – this is the sum of squares for the 1st h residual correlations, i.e. $Q_h = n(n+2)\sum_{k=1}^{h} \hat{\rho}_z^2\ (k)/(n-k)$ in the univariate case, which has an asymptotic chi-squared distribution with $h-p-q$ degrees of freedom. If $Q_h > \chi^2_{1-\alpha_C}(h-p-q)$ the adequacy of the model is rejected at level α_C – typically 0.05. In the multivariate case, the statistic has a chi-squared distribution with $m^2(h-p-q)$ degrees of freedom and is

$$Q_h = n^2 \sum_{k=1}^{h} Tr(\hat{\Gamma}'(k)\hat{\Gamma}^{-1}(0)\hat{\Gamma}(k)\,\hat{\Gamma}^{-1}(k))/(n-k). \tag{3.64}$$

3.4. The Wavelet Transform

Mathematical analysis of a time series in the time domain yields no direct information about the distribution of spectral components, while Fourier analysis in the frequency domain involves complete loss of specific information about the time domain. In this respect, it could be said that the Fourier transform goes 'too far' by eliminating all time resolution in lieu of frequency resolution, particularly if it is suspected that the dynamics of the series may change over time. One solution to this is the short-time Fourier transform, which takes a sliding window across a time series and calculates the Fourier transform of the series inside the window. In this case, the analyst has a fixed amount of resolution in both the frequency and time domains.

Heisenberg's uncertainty principle, originating from quantum mechanics, may be usefully applied to signal processing. In this new context, it states that one cannot simultaneously achieve any desired level of resolution for frequency and time since a signal cannot have, simultaneously, a precise location in time and precise frequency. A system of analysis which actively works with this principle, to obtain the maximum amount of information about a time series (signal), is wavelet transforms. It provides high frequency resolution on long time-scales and accurate time resolution for high frequency events, i.e. able to capture features that are local in both time and frequency. Wavelet methods provide a natural platform to deal with the time varying characteristics found in many real-world time series, avoiding the need to assume stationarity, and with features such as structural breaks and volatility clusters made clear.

This section provides a very brief introduction to wavelet analysis, and its potential use in this project – largely based upon the work of Gençay, Selçuk and Whitcher [112] and Strichartz [113]. As an example, a function $x(t)$ on the interval [0,1] may be expanded as a Fourier series:

$$x(t) \;=\; b_0 + \sum_{k=1}^{\infty} b_k cos(2\pi kt) \;+\; a_k sin(2\pi kt)\,, \tag{3.65}$$

or also as a Haar (square pulse) function series:

$$x(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^{j-1}} c_{jk}\, \psi(2^j t - k), \quad \text{where} \tag{3.66}$$

$$\psi(t) = 1 \text{ for } 0 \le t < 1/2, \quad \psi(t) = -1 \text{ for } 1/2 \le t < 1, \text{and } \psi(t) = 0 \text{ otherwise.} \tag{3.67}$$

This is an example of a discrete wavelet expansion, and $\psi(t)$ is known as a mother wavelet, which is both stretched and shifted to create the original function. A wavelet can be any function that obeys a basic rule, known as the wavelet admissibility condition

$$\int_0^\infty (|\Psi(\omega)|/\omega)\ \mathrm{d}\omega < \infty\,, \tag{3.68}$$

where ω is angular frequency and $\Psi(\omega)$ is the Fourier transform of $\psi(t)$. The condition implies that the function must be zero mean and have unit energy. As the name suggests, wavelets can typically be visualized as a brief oscillation, and are purposefully crafted to have specific properties that make them useful for signal processing.

The continuous wavelet transform (CWT) is a function of two variables $W(u_{WT}, s_{WT})$ and is obtained by simply projecting the function of interest $x(t)$ onto ψ via

$$W(u_{WT}, s_{WT}) = \int_{-\infty}^{\infty} x(t)\,\psi_{u,s}(t)dt \tag{3.69}$$

$$\text{where} \quad \psi_{u,s}(t) = \psi((t - u_{WT})/s_{WT})/s_{WT}^{1/2}\,. \tag{3.70}$$

The reverse transformation may be used to reconstruct the original function.

Since the CWT is a function of two arguments, while the original has only one, it contains a large amount of redundant information. It is therefore possible to reduce the task of the wavelet transform from one involving continuous parameters to a set of samples taken from the CWT, a form of discretisation, leading to the discrete wavelet transform, DWT. The sampling points are determined by limiting scale to $s = 2^{-j}$ for all integers j, and translations are limited to $u = k.2^{-j}$ for all integers k. The corresponding wavelets

$$\psi_{j,k}(t) \;=\; 2^{j/2}\,\psi(2^j t - k) \tag{3.71}$$

form an orthogonal basis. If instead we chose $u = k$, the resulting transform is known as the maximum overlap DWT, or MODWT.

Wavelet analysis is very useful in the fitting of long memory models, and also for simulating them [114]. Applying wavelet transforms to long memory processes, Jensen [115] established a log-linear relationship between the wavelet coefficients' variance and the scaling parameter, equal to the long memory parameter. This forms the basis of a consistent OLS estimator for the parameter. Later, Jensen [116] used the relationship as the basis for an alternative maximum likelihood estimator for the differencing parameter, d, of an ARFIMA (p, d, q) long memory process that is invariant to unknown means, model specification, and contamination. Jensen showed that since the wavelet coefficient covariance matrix is sparse, and can be replaced with a diagonal one to a good degree of accuracy, the computational burden of the estimator is less than those associated with the exact MLE method.

Whitcher [117] used the same log-linear relationship in order to estimate the differencing parameter for Gagenbauer processes, developing both an ordinary least squares

estimator and an approximate maximum likelihood estimator, making use again of the sparsity of the covariance matrix. One difference to Jensen is that Whitcher uses MODWTs in addition to DWTs. As these examples demonstrate, wavelet coefficient covariance, sometimes simply referred to as wavelet variance, is the fundamental tool in wavelet analysis, analogous to periodograms in spectral analysis.

Gençay, Selçuk and Whitcher [118] state that conventional time series analysis, which focuses exclusively on a time series at a given scale, may lack the ability to explain the nature of the 'data generating process'. An econometric model that successfully explains daily price changes, for example, is unable to characterize the nature of hourly price changes. Further, statistical properties of monthly price changes are often not fully covered by a model based on daily price changes, they state. This may also be the case for wind, the two relevant scales being (i) hour-to-hour changes that occur within synoptic types, and (ii) the procession of those types.

The authors simultaneously modelled regimes of volatilities at multiple time scales through wavelet-domain hidden Markov models (HMMs), distinct from traditional HMMs as they capture dependencies in the two-dimensional time-frequency plane. They established an important stylized property of volatility across different time scales, named asymmetric vertical dependence. It is asymmetric in the sense that a low volatility state, or regime, at a long time resolution is most likely followed by low volatility states at shorter time resolutions. On the other hand, a high volatility state at long time resolutions does not necessarily imply a high volatility state at shorter time resolutions.

Wavelet methods are also powerful in the analysis of multivariate processes, as wavelet cross-covariance decomposes the cross-covariance between two time series on a scale-by-scale basis, revealing how the association between them changes as a function of time scale. Within the framework of long memory multivariate processes, fractal connectivity is a particular model in which the low frequencies (coarse scales) of the cross-spectra of each variable pair are determined by their auto-spectra. This is the case when the long memories in each of the components arise from the same single mechanism – surely true for wind. Wendt et al. [119] developed a statistical procedure for testing for fractal connectivity amongst data, anchored naturally in a wavelet framework.

Analysis in a wavelet framework is considerably more challenging than in purely the time or frequency domains – precisely because more information is yielded about the

dynamics of the process, particularly in the multivariate case. Indeed, for a 20-dimensional process with timescales of interest ranging from hours to decades, wavelet analysis seems unfeasibly complicated. One way to ease the problem, to an extent, is to transform the dataset into its principal components (as described in the following paragraph) and ignore those which make small contributions to the total variance. In this way the dimensionality of the problem may be reduce, perhaps by half, without much loss of accuracy.

The PCA procedure is to calculate the eigenvectors of the series' covariance matrix, then create a matrix with rows that are those vectors transposed, and arranged in decreasing order of their associated eigenvalues. With the series arranged into a matrix where each row is a dimension and with time 'flowing' from left to right, this matrix must be pre-multiplied with the eigenvector matrix to transform the series into a set of independent modes of variability. The first row of the new matrix is the mode which contributes the most variance to the total, and they are arranged in decreasing order. The first rows constitute the principal components of the multivariate series.

While the components are independent, in the sense that the expectation values of their covariances are zero, the fact that wind is a nonlinear, non-stationary process means that some relationships between the components would probably remain. These relationships would probably be different of different timescales, and multivariate wavelet analysis could capture them. There is however a major drawback of modelling the principle components, rather than the series directly: their individual dynamics are likely to be quite different. The 1st component, for example, is likely to correspond roughly to a spatial average across GB and will be dominated by lower frequency variability. The 5th and 6th components, in contrast, are likely to be more spatially localised (i.e. consist of much heavier contributions from certain zones than others) and will be dominated by higher frequency variability. As such, it would be difficult to construct a multivariate model where each zone has the same structure, but different coefficient values. It may also be difficult to construct any kind of closed-form expression that captures the inevitable relationships between components, as is discussed further in the next section.

3.5. Non-Gaussian Multivariate Processes and Copula Functions

The following discussion re-introduces the assumption that the multivariate process to be modelled is stationary. For multivariate statistical models, the dependencies between non-normally distributed components can cause problems. While there are many kinds of parametric univariate probability distributions, only in a few cases is there a natural multivariate analogue, and finding the best choice is by no means a trivial problem. The cross correlation of wind speeds may be quite different during calm periods compared to stormy periods, for example – an issue not discussed in section 3.1.5. One method used to circumvent this problem is multivariate Gaussian mixture models, but such an approach has two serious problems: they become complicated for higher dimensions, and do not account for the original marginal distributions.

It is therefore more useful to work with copula functions, also simply called cupulas. Indeed, they are an area of growing application in meteorology, as described by Scholzel and Friederichs [28]. Copulas are mathematical reformulations of multivariate CDFs, while multivariate PDFs may be reformulated in terms of copula densities and marginal PDFs.

Considering a m-dimensional random vector $\underline{X}$ with marginal CDFs $F_{X_1}(x)$, . . . , $F_{X_m}(x)$, Sklar's theorem (1959) states that the joint distribution $F_{\underline{X}}$ of this vector can be written as a function of its marginal distributions, i.e.

$$F_{\underline{X}}(x) = C_{\underline{X}}(F_{X1}(x_1), \ldots, F_{Xm}(x_m)). \tag{3.72}$$

Assuming continuous and differentiable distribution functions $u_i = F_{X_i}(x)$, we can express $C_{\underline{X}}$ as

$$C_{\underline{X}}(u_1, \ldots, u_m) = \int_0^{u_1} \ldots \int_0^{u_m} c_{\underline{X}}(u'_1, \ldots, u'_m) du'_1 \ldots du'_m. \tag{3.73}$$

The function $C_{\underline{X}}$ is called a copula and $c_{\underline{X}}$ the corresponding copula density. The important consequence of Sklar's theorem is that every joint probability density can be written as the product of the marginal probability densities and the copula density:

$$f_{\underline{X}}(\underline{x}) = f_{X_1}(x_1) \ldots f_{X_m}(x_m)\, c_{\underline{X}}(u_1, \ldots, u_m). \tag{3.74}$$

The copula density is equal to one for independent random variables, but for dependent variables the question remains of how to formulate and estimate it. In fact, there is no general or canonical way to formulate the copula and to assess the relationship among the random variables. However, parametric copula functions exist, grouped into families, as do empirical methods for calculating which family best fits a set of observed time series. Two important families are the elliptical and the Archimedian copulas. Examples of Archimedean

copulas are the Clayton, for which correlations are stronger for smaller values, and Clayton survival, where the opposite is true. The Gaussian copula simply captures the relationship between MVN variables.

Despite the existence of such standard functions, the problem of finding parametric distributions for high dimensional random vectors remains complex, and the number of parameters very large.

3.6. Markov Chains

This section elaborates upon the brief introduction given in Chapter 1 to discrete Markov models, the main rival to regression-based time series models.

Discrete Markov models describe stochastic processes where the observed variable of interest is modelled as being in one of a finite set of states, often based upon division of the range of observed values into bands. States are often chosen such that they have similar probabilities of occurring, and top and bottom states may be open intervals when the support for the variable is infinite. The state of the variable at time t is described by a probability distribution represented by the vector $\underline{\Pi}_t$. If there are s_M states then the vector has s_M dimensions, with the j^{th} element representing the probability that the variable is in state j at time t. Being a Markov process, the state evolution from time t to $t+1$ is governed by

$$\underline{\Pi}_{t+1} = \boldsymbol{Q}\underline{\Pi}_t \tag{3.75}$$

where the elements q_{ij} of the transition matrix $\boldsymbol{Q}$ give the probability that the variable will make a transition from state i to state j. The assumption here is that the state at time $t+1$ is dependent only on the state at time t, satisfying the Markov condition. Such processes are therefore discrete in two senses: the states of the variable and units of time.

Markov chain Monte Carlo simulations are a means of simulating a discrete Markov process with a defined transition matrix, whereby the simulated variable X_t is assigned to a specific state with 100% probability for each time step, based on $\underline{\Pi}_t$ and an uniform random deviate produced for each time step. This process involves extracting a CDF for each hour from $\underline{\Pi}_t$ – i.e. the range [0,1] is divided into segments associated with each state, and the uniform deviate generated will fall within one segment, determining the state. It is possible to generate an additional random number for each time step to determine, based on the reverse cumulative density, exactly where the variable lies within the range of the state.

For ARMA and similar extension models, innovations X_t - $\hat{X}_t$ are not conditional on X_t, which is clearly unrealistic. For example, when simulating an ARMA process and X_t has an extremely high value, the probability of the next innovation taking a large positive value is treated as identical to when X_t is small. This is clearly not the case in reality, and the problem does not apply to Markov Chains, giving them a significant advantage.

A disadvantage of Markov chains is that the Markov condition means that persistence may be under-represented for lags greater than a few time-steps. A solution is that Markov chains may be extended to higher orders, in which the future (i.e. $t+1$) state is assumed to

depend on a few past states in addition to the current (time t) state. For a model of order l there is a dependency on $l-1$ past values. The memory-less quality has not been abandoned, the concept is rather that the 'current' situation is described by a sliding set of l states – current and past values. In general a model of order l with s_M states requires a s_M^l x s_M transition matrix with s_M^{l+1} parameters. This is because there are s_M^l 'present' states for times t to $t-l$, transitioning to s_M possible states at time $t+1$. Approximate models also exist, such as one defined by

$$\underline{\Pi}_{t+l+1} = \sum_{i=1}^{l} v_i \boldsymbol{Q}_i \underline{\Pi}_{t+l+1-i}, \quad \text{with} \quad \sum_{i=1}^{l} v_i = 1. \tag{3.76}$$

Depending on the desired level of sophistication for the model, the transition matrices may be identical, requiring a total of $s_M^2 + l$ parameters, or different, requiring $l(s_M^2 + 1)$ parameters.

Another important extension is the multivariate discrete Markov model. We consider, for simplicity, only 1st order models with m observable variables which may individually be in one of s states. The state of the system is now a combination of individual states, so there are s_M^m states in total, requiring a s_M^m x s_M^m matrix and s_M^{2m} parameters. Such a model can capture complex joint distributions without difficulty, but the price to pay is a very large transition matrix, unless the number of dimensions is very small.

The matrix size can be reduced through the use of an approximate model considering all relationships between pairs of variables only, with the defining equation still 3.75, but where

$$\underline{\Pi}_t = [\underline{\Pi}_t^{(1)'}, \dots, \underline{\Pi}_t^{(m)'}], \tag{3.77}$$

and $\underline{\Pi}_t^{(j)}$ is the state probability vector for the j^{th} variable, and

$$\boldsymbol{Q} = \begin{bmatrix} v_{1,1}\,\boldsymbol{Q}^{(1,1)} & \cdots & v_{1,\mathrm{m}}\,\boldsymbol{Q}^{(1,\mathrm{m})} \\ \vdots & \ddots & \vdots \\ v_{\mathrm{m},1}\,\boldsymbol{Q}^{(\mathrm{m},1)} & \cdots & v_{\mathrm{m},\mathrm{m}}\,\boldsymbol{Q}^{(\mathrm{m},\mathrm{m})} \end{bmatrix}, \tag{3.78}$$

Where $\boldsymbol{Q}^{(j,k)}$ is the transition probability matrix relating states in the k^{th} series to the states in the j^{th} series, and the $v_{i,j}$ are weighting factors. Such a model requires $m^2(s_M^2+1)$ coefficients, which represents a significant reduction unless m and s_M are small. To simulate such a process, an uniformly distributed random number is required for each variable to decide their states at time $t+1$. This set of numbers will have a highly complex dependence structure which can in principle be calculated using the copula approach, but would be extremely challenging with m = 20, as is the case for the process to be modelled in the current research.

3.7. Chapter Summary

The chapter began by presenting the basic principles and assumptions behind ARMA models. The models were presented as finite and infinite response filters with white noise inputs, and the resulting spectra were discussed. Many model extensions were presented, including seasonal models and those requiring differencing and/or seasonal differencing. Fractional differencing operators were introduced as a means of representing long memory processes – both ordinary and Gagenbauer. Since it was reported in Chapter 2 that wind speed series display conditional heteroskedasticity, a variety of models were presented that can capture such behaviour. Multivariate models of conditional variance were presented, but it was seen that even simplified ones imply an undesirably large number of parameters for 20 dimensions.

The Box-Cox transformation was introduced along with other, more sophisticated methods of transforming a set of series into an approximate MVN – but these are unsuitable when the number of dimensions is so high. For the fitting of models to (transformed) finite samples, it was shown that correlograms are the fundamental tool in the time domain, while periodograms have the same role in the frequency domain. Various model fitting methods were presented, including the Yule-Walker Equations, the univariate and multivariate Levinson-Durbin algorithms and the approximate but simple Hannan-Rissanen procedure. The maximum likelihood principle was presented, along with expressions for the likelihood function. Emphasis was placed on the complexity of such expressions, particularly for multivariate mixed ARMA models with several extensions.

Wavelet transforms were presented as a means of analysis that is superior to purely time of frequency based analysis due to the methodology's utilisation of Heisenberg's uncertainty principle, applied to signals. The distinct advantages of the wavelet framework are that stationarity need not be assumed, and that relationships between dimensions can be analysed separately on different timescales. The methodology is however more challenging, particularly when working with 20 dimensions and a large range of timescales. The merits and problems associated with reducing the number of dimensions through Principal Component Analysis were presented. Copula functions were presented as a means of dealing with multivariate processes that are assumed to be stationary, but with nonlinear joint distributions. The difficulty of finding the correct copula function for a dataset was discussed.

The Markov chain simulation methodology was presented, along with its natural advantage – inbuilt conditionality, for both marginal and joint distributions. Multivariate and higher-order Markov models were presented, but it was seen that even highly simplified versions require very large matrices when the dimensionality is high.

Chapter 4

Previous Wind Modelling Work

This chapter reviews previous wind modelling work found in the literature. The detailed picture of the GB wind resource presented in Chapter 2 will be used to critically evaluate the extent to which the reviewed models are capable of reproducing the resource's behaviour, in all its complexity. This chapter not only reviews the models but also, where appropriate, the model fitting and validation methodologies, facilitated by the presentation of relevant mathematical background in Chapter 3. Some of the literature reviewed was also included in Chapter 2, but where the focus previously was on insight regarding the nature of the resource, here the focus is on methods developed for reproducing the reported behaviour and properties.

4.1. Previous Modelling Conducted at the University of Bath

4.1.1. The Bath Wind Model

The Bath Wind Model was developed by power systems researchers at the University of Bath, notably Dr. Rod Dunn (the supervisor for this research), Dr Marcos Miranda and Dr Furong Li. Their work is described here as documented in a published conference paper [25] and a study commissioned by National Grid [27].

As previously stated, the Bath wind model treats wind as a 4^{th} order VAR process. The methodology divides GB into 20 zones, referred to here as the Bath Zones, and power system simulations using the Bath Model involve calculating the balance of total available generation and demand within each zone for each trial, leading to LOLE calculations (or, in principle, any risk metric).

The Bath Zones are based on 17 study zones adopted by National Grid, as presented in their Seven Year Statement [26], with 3 additional zones in Scotland, as figure 1.1 shows. The zones represent areas of the country with strong internal electrical connections but with weaker interconnections to the rest of the system. The Bath Zones in England and Wales are

identical to those of the Seven Year Statement, while Scotland's three additional zones were created and the boundaries of existing ones slightly changed in response to Scotland's disproportional importance in terms of future wind capacity. The model was fitted to historical wind speed data collected at a Met Office Station within each zone. The stations were chosen based on them offering the longest consistent historical measurements - the result being data beginning in 1983/84 and going on to at least 1994 for a few stations, and up to the year 2000 for some others.

Model parameters were estimated by using Matlab's System Identification Toolbox [122], which favours a method based on the Yule-Walker equations. No reason was given for the choice of AR(4) as the model order. Matlab offers another toolbox suitable for the fitting of ARMA models – Econometrics [123], but neither allows the fitting of mixed or MA models in the multivariate case. The Bath Model involved no pre-treatment of the data, such as Box-Cox transformation or the removal of trends, cycles or seasonal effects. This is partially justifiable since only data from winter months (Dec - Feb) was used, in order to reflect the resource during peak times, and the diurnal seasonality is subtle during these months. It appears that wind speeds were modelled as being normally distributed – clearly not very accurate, but as Chapter 2 reported, more accurate during winter months. With the model fitted, Matlab was used to simulate synthetic time-series of any desired length. While temporal correlations ensure that the output time series represent realistic wind speed sequences, the simulation does not associate any time of day or day of year with the wind speeds.

The next stage in the production of wind powers is to scale-up wind speeds from those experienced at a height of 10m to those expected at the assumed turbine hub height of 80m. In [25], Miranda and Dunn achieved this using the power law discussed in Chapter 2, section 2.2, using a constant α value of 1.7. These wind speeds are then transformed into normalised wind power outputs using a manufacturer's wind power curve for the 3MW Vestas V90. These are converted into zonal power outputs based on a specific spatial arrangement of wind capacity. With all turbines assumed to have capacities of 3MW, the number of turbines in a zone is given by capacity/3MW. The availability of each turbine for each trial is modelled as an independent Bernoulli variable, with assumed availabilities of 97% for onshore and 92% for offshore (sampling from zonal binomial distributions would be easier). There is clearly an implicit assumption that offshore wind-farms are located close to the mainland, as they are treated as experiencing the same wind conditions as their associated mainland zone.

A more sophisticated version of the later stages of the methodology was presented in the National Grid commissioned study [27]. Some aspects of this study were sub-contracted to the wind energy consultancy Garrad Hassan Ltd (GH). The first difference in methodology was the introduction of four different types of windfarm location:

1. Coastal – within 5km of coast and elevation under 300m AMSL (Above Mean Sea Level);
2. Lowland – greater than 5km from coast and elevation under 300m AMSL;
3. Upland – elevation above 300m AMSL, any distance from coast; and
4. Offshore.

In order to be transformed into wind power outputs, the generation scenario had to specify the generation capacity of each type in each zone. Clearly not all zones have each type, particularly the land-locked ones. The methodology also differentiates between transmission and distribution connected capacity for each location type within each zone, to account for transport losses. The scenario must therefore provide capacities for both types.

The next additional sophistication is taking account of the fact that the wind climates at wind-farm locations differ from those at the Met Office stations. This is achieved through constant speed up ratios (for constant heights of 10m) – a different one for each location type in each zone. These values were derived by Garrad Hassan based on operational and meteorological data it has collected and UK NOABL, a wind resource database [124]. These are mostly greater than unity, but range from 0.6 for lowland sites in zone 4 to 1.81 for offshore sites in zone 11. Different 10m to hub-height speed-up ratios are also introduced for each location type, based on similar data, but in this case they do not vary between zones. They are: coastal – 1.34, lowland – 1.48, upland – 1.2, offshore – 1.25. These are all smaller than the value 1.7 chosen for the simpler methodology, reflecting the lack of 10m – 10m speedups in that case.

The last difference is more sophistication with regard to the turbine power curves. Rather than a curve published by a manufacturer, conversion tables were developed by GH using their ‘Bladed’ software. Two power curves were in fact provided – representing high and low speed turbines. The latter represents a slightly larger turbine with its maximum output capped, allowing it to achieve the rated output at a lower speed, 11m/s rather than 13m/s. The GH curves implicitly account for realistic losses, including wake and electrical losses. These are consistent with a thorough exploration of the difference between the manufacturer’s curve and a realistic one for a wind farm, based on measurements, is provided by Hayes et al.

[125]. The assumed availability for onshore turbines is again 97% but a slightly smaller 90% is assumed for offshore. Offshore turbines are again 3MW, but onshore turbines are assigned the smaller value of 2MW.

4.1.2. Validation of the Bath Wind Model

Validation of the Bath model in [27] began with an examination of its ability to reproduce the basic statistics of the historical wind data. Comparisons are made of the model output and input data's mean wind speed and variance for each zone, and the plotted results show very good agreement. This is in fact a trivial result since the model's means and variances have been set as exactly equal to their sample value for the historical data. Spatial correlation was verified by plotting the rows of the correlation matrix for concurrent wind speeds, for both data and model, and the match is seen to be excellent. Again this is not surprising, since the auto-regressive coefficients are derived from correlations found in the data, although validation was required with respect to this characteristic. The relationship between these correlations and distance was explored, and found to be roughly consistent with the $exp(-d_p)$ relationship discussed in Chapter 2 (d_p is the distance between pairs). This relationship was shown in figure 2.2, where the number of pairs is much greater. Although this relationship is worth noting, it will not form part of the model fitting process.

Validation of the wind speed modelling is followed by validation of the wind power output, beginning with comparisons between winter load factors derived from model output data and real wind-farms in the same zones, presumably provided by GH. It was found that the model load factors are higher - in some cases only by a few %, in others by up to 32%. The main explanation proposed in the report is that the assumed turbine availabilities are too high. This is true to some extent – a recent study published by National Grid [57], for example, used availabilities of 95% and 85% for onshore and offshore turbines, respectively. However, another reason may be that the historical data are not Gaussian, while the simulated series are, so that the medians of the simulated series are too high.

A somewhat stronger validation of the Miranda and Dunn methodology can be found in [25]. Here the simpler wind speed model is fed into a (fairly simple) Monte-Carlo simulation of the GB system in order to investigate the LOLE of various total installed capacities. The loss of load probability (LOLP) is discussed, but defined as the number of winters per 100 years during which a forced loss of load event can be expected. As discussed in Chapter 1, this is not the most rigorous variant of the LOLP concept, as it gives no indication of the number of hours

during which generation is unsufficient. Nevertheless, an event occurring in 9 out of each 100 winters is a well-established target from the time of centrally planned systems, and this level is found to occur when a plant margin of around 22% is assumed in the simulations. These two values are consistent with the current market practice, suggesting that the methodology is realistic in some important aspects, leading the paper's authors to conclude that the method can be confidently applied to more radical future scenarios where the installed capacity of wind generation is higher.

4.1.3. Other Wind Modelling Work Conducted at the University of Bath

Two papers written by Drs Miranda and Dunn, with assistance from statistician Dr. Gavin Shaddick, explore Bayesian approaches to GB wind resource characterisation and forecasting. The first [126] is concerned with one-hour-ahead prediction at a single site, by modelling wind speeds as an AR(6) process. The historical data used was again Met Office records, and the example site chosen is Lerwick in the Shetland Islands. The Bayesian method of Markov Chain Monte Carlo simulation is used to obtain probability distributions for the parameters, the mean of which may be compared with the results of frequentist approaches.

It is assumed that wind speeds are Weibull distributed, and thus the data must be transformed to have a normal distribution, using a Box-Cox transformation, such that normal priors and likelihoods may be assumed in the Bayesian analysis. A likelihood function is presented for the Box-Cox parameter λ_{BC}, and a plot shows that its maximum value is reached for λ_{BC} = 0.5. This method was compared with a frequentist approach, using Matlab's 'boxcox' function, that gave a very similar value.

A table presents values for the six autoregressive coefficients, one set derived using Bayesian sampling, another using Matlab's System Identification Toolbox. Both methods are in agreement that the first (i.e. $t-1$) coefficient is slightly greater than unity, while the others are all very small, mostly negative numbers. Using Matlab to find the roots of the AR polynomial for both parameter sets shows that the smallest root is 1.0714 in both cases (i.e. the values are identical to 4 decimal places). The process clearly has strong persistence and is very close to requiring 1st order differencing.

Having the $t-1$ parameters close to unity indicates a strong resemblance to the persistence model, the simplest possible time series model, which states that the best 1-step-ahead forecast for a process is its present value. The persistence model is commonly used as a

benchmark against which the predictive power of a more complex model can be compared. The R^2 value for the Bayesian model – i.e. percentage of variance 'explained' by the model, was 83.4% compared to 82.7% for the persistence model. This suggests that the autoregressive model's predictive power would not be significantly compromised by reducing the model order considerably. Despite this, the authors state that lower order models may fail to capture periodical variations of the wind speed. It is also stated that predictions could be improved considerably through the use of more complex models, such as autoregressive moving average (ARMA), given that such models can "characterise seasonality and other level variation effects in the data".

Given that mixed ARMA models may be expressed as AR(∞) models, they may well offer a more parsimonious means of capturing behavioural subtleties that low order AR models cannot, and this seems particularly true in the multivariate case. However it is clear that extending from pure AR to mixed models cannot capture specific seasonal behaviour, as Miranda, Dunn and Shaddick appear to suggest. Justification for adopting a Bayesian framework, where the calculation of coefficients is more involved, is said to be the framework's flexibility - allowing the inclusion of expert knowledge e.g. on effect of atmospheric pressure to improve predictions. It seems though that for parameter estimation, the additional computational expense is not justified.

Their second paper exploring the Bayesian framework [127] extends the analysis to the multivariate case. Rather than short term prediction, the goal here is to demonstrate a methodology for the characterisation of wind speed at a site for which no data is available. It uses the Bayesian approach to model the spatial correlation between different sites in the region, for which data is available, and infer wind speeds for the new location – in effect creating a virtual weather station. The datasets used in the model were rather short - one year (1997) of hourly average wind speeds from four weather stations at the north of Scotland. Predictions are made for a 5th location in the area, and compared to actual data obtained for this site to validate the methodology. Most details of the methodology are not relevant here, but several points within the paper are worthy of discussion.

Data was missing from each dataset, with the percentage ranging from 1.39% to 7.91%. While the methodology used allows the inclusion of these points as missing data, calculation of correlation coefficients included only hours for which data was present for every location. This amounted to 87% of the hours, a reasonably high number, although if 20 locations were used as in the Bath Wind Model, one may expect this percentage to be

unacceptably low. Box-Cox transformation is again deemed necessary, and the best choice for λ_{BC} is investigated as previously for each site individually. It is stated that for consistency, each site must be transformed using the same value, and the best compromise was deemed to be 0.3 – significantly different from the previous value. The methodology assumes that the wind resource is stationary and does not include any seasonal component. When analysing model residuals – i.e. the unstructured noise term, no noticeable trend or cycle was found, rather surprisingly. Clearly the presence of seasonalities known to be present is difficult to detect in simple model residuals, for reasons that are not clear.

4.2. Other ARMA Models for Wind Speed

4.2.1. Univariate ARMA Models

Monbet et al. produced a comprehensive report, *Survey of stochastic models for wind and sea state time series* [97], which states that the obvious non-stationarities in wind speed time series are generally dealt with in two ways. The first approach is to build a model of the form

$$X_t = m(t) + \sigma(t)X_t^{Stat}. \tag{4.1}$$

Here $m(t)$ and $\sigma(t)$ are deterministic periodic functions with period one year, and also possibly the diurnal period, and $\{X_t^{Stat}\}$ is a stationary process. The functions $m(t)$ and $\sigma(t)$ may be estimated non-parametrically or modelled as a sum of parametric functions such as cosine and sine.

The other approach is to suppose that the process is piecewise stationary and to fit separate models for each month or each season of the year. This is essentially the Bath Wind Model approach, which has assumed that wind is stationary over the winter. This method is reliant on seasonal change being negligible during the month or season, which may not be entirely accurate, and it introduces artificial ruptures between successive periods if modelling the entire year. The survey provides examples in which wind speed is modelled as AR(1), AR(2) and AR(4) processes, with more complex models reported as providing no significant improvement.

Daniel & Chen [128] are concerned with fitting ARMA models to hourly averaged wind time series from Jamaica – fitting separate models for each month. Their methodology is to: (i) make distributions Gaussian; (ii) eliminate diurnal trends to make the process zero mean and with constant variance; (iii) establish model orders; and (iv) fit parameters.

Step (i) makes use of the fact that the Weibull distribution with k_W = 3.6 is very similar to the Gaussian distribution, and that if the random wind speed U_t is Weibull distributed with shape and scale parameters k_W and C_W, respectively, then $U_t^{\delta_W}$ has a Weibull distribution with parameters k/δ_W and C^{δ_W}. So, to convert to nearly Gaussian one must raise the wind speeds to the power $\delta_W = k_W/3.6$, in a process which is a simplification of the Box-Cox transformation. Translation of wind speeds to a new location with e.g. increased occurrence of extreme winds beyond a simple speeding up of all hours means giving the distribution a fatter tail, which means changing the shape parameter to k_W'. This can be achieved by raising value to the power $3.6/k_W$' (rather than $3.6/k_W$) in the reverse transformation, and then finding the

appropriate value for C_W. Unfortunately, this is not possible for the current research project, since shape parameters for realistic heights at realistic wind turbine locations are not available

Returning to [128], both information criteria were used for model order selection. The more parsimonious BIC gave AR(1) as the best model order, while the AIC chose AR(3), leading to a compromise choice of AR(2). The method of least squares is used for parameter estimation and the portmanteau statistic used for diagnostic checking. This statistic is described in section 3.35 and is given by equation 3.64.

Desrochers et al. [129] develop a method for determining the cost-effectiveness of wind energy and the economic limitations of penetration into electrical power systems. For the wind speed one year's hourly data, divided on a monthly basis, was studied for three areas in Canada. The best ARMA model for each month and each area was selected and, among these, the three model types that occurred most often were selected for use in a simulation: AR(2), AR(3) and ARMA(1,1). The procedure was then to fit one of these models separately to each month and location, trying each type first and selecting the best based upon a portmanteau statistic and the principle of parsimony. A power curve was used to convert to wind power outputs.

A paper by Billinton, Chen and Ghajar [131] presents two different univariate ARMA models. Only the 1st is of interest here, takes the form of equation (4.1) and was fitted to hourly averaged wind speed data obtained from a government agency. The authors state that it has been shown that any stationary stochastic system can be approximated as closely as required by an ARMA$(p, p-1)$ model, so that the question of determining (p, q) becomes that of determining p.

Estimation of the deseasonalised ARMA model parameters was achieved with the use of non-linear least square optimisation. Nonlinearity here refers to the fact that innovations had to be implied for an assumed full set of parameters, so the quality of the starting values plays a very important role in the convergence of the iterations. The Gauss-Newton method with the halving mechanism was used to minimize the sum of squares, which is a strategic modification of the classic Gauss-Newton method. Since this method encounters difficulties in some instances, the Marquart procedure is also used to improve the convergence – this is useful to know for the selection of optimisation algorithms in *Matlab*'s optimisation toolbox later.

To check the adequacy of the proposed models, a procedure described as the F-criterion test was used. This methodology is based on a metric involving the improvement in the residual sum of squares for a model of order $(p+1, p)$, compared to a model of order $(p, p-1)$. If this metric is greater than the F-distribution used in hypothesis testing, with suitable degrees of freedom and confidence level, then the improvement in the residual sum of squares is significant at that level and therefore there is evidence that the ARMA(p,$p-1$) model is inadequate. The method, along with several of the statistical tests on model residuals described in Chapter 3 lead to the conclusion that ARMA(3,2) is the best choice. The methodology also established that if a pure AR model is required, the same level of accuracy is achieved by an AR(8) model – considerably higher than the conclusion of other researchers described above, due to using a rigorous but more prescriptive criterion.

The autocorrelogram for synthetic data generated with this model showed very good agreement with that for real data, although the plot only goes up to a 32 hour lag. Plots of the annual seasonality and the diurnal pattern for an example month also displayed excellent agreement. The paper shows that the sample auto-correlation functions of different years may be significantly different, thus a wind speed model based on only one year of actual wind data should be used with caution. It seems unlikely that this aspect of inter-annual variability can be reproduced accurately by this model.

4.2.2. Multivariate ARMA Models

A multivariate wind speed model very similar to the Bath wind model has been developed at the University of Strathclyde, as reported by Hill et al. [132]. One major difference is that the Strathclyde model not only appropriately models spatial correlations across large areas but also takes due account of seasonal and diurnal variations, and how those diurnal variations change with season. Their concern is with the modelling of wind speed at 10m, specifically the extent to which synthetic wind speed datasets are statistically similar to the set of hourly-averaged Met Office historical datasets on which the model was trained. Two model types were developed – univariate ones for each zone, and one multivariate. Hill et al. state that in further work, winds speeds will be converted to normalised power outputs in the same way as the Bath model, whilst acknowledging the complexity of this relationship, due to changing thermal stratification and shadowing effects.

The University of Strathclyde model is, like the Bath model, based upon the 17 zones defined by the National Grid for their Seven Year Statement [26]. The Strathclyde authors found, however that it was not always possible to find a station close to existing wind farms in each SYS region with good quality data and so 14 data sites were chosen to cover the U.K., with the datasets ranging from 14 to 28 years in length.

The authors treated all seasonal patterns as entirely deterministic in nature, finding that the annual component is well modelled by a low order Fourier series. The diurnal component was found to be more complex, with a need to define 4 separate seasonal patterns. Should the year be broken into more individually shorter time periods to improve resolution, they argue, the quality of the trend identified in any period would decline due to reduced data. A fairly consistent picture emerged across a number of sites and low order Fourier series models were again fitted. The R^2 value for the fitted annual seasonality was found to be 0.948, while it was greater than 0.99 with the diurnal patterns (slightly smaller for winter).

Hill et al. claim that de-trending to remove the seasonal and diurnal variations has the added advantage of making the probability distribution of the long term wind speed approximately Gaussian, so that no further transformation is required. A diagram of the wind speed distribution at one location, before and after detrending, is provided as evidence for this argument. Indeed it shows that the asymmetry associated with the Weibull distribution is much reduced, but it is not entirely eliminated. Small peaks at 0m/s and 2m/s in the original distribution are eliminated, and the distribution is significantly smoother generally.

For model order selection on the detrended series, the authors inspected the ACF and PACF (correlation functions) for the datasets. These are presented in the paper up to a lag of 48 hours, for one location. It is stated that no MA terms are necessary in the model since the ACF smoothly decays to zero – this is not exactly true, rather it decays to a small positive value. It is noted that the PACF indicates the presence of two or possibly three AR terms, as it is significantly different from zero at these lags. A more detailed analysis involving minimization of the one-step-ahead prediction error showed that varying the number of parameters from 2 to 4 produced less than 1% improvements. A parsimonious AR(2) model was selected, and similarly for the VAR.

The univariate AR coefficients were estimated initially using OLS, then the Yule-Walker approach was used and the estimates obtained taken to be final. For the multivariate model

OLS based techniques were used, as discussed in the VARMA model fitting literature. The ACF of the residuals, for both univariate and multivariate models, were inspected to check that they approximate to white noise. The authors state that figures provided show that this is indeed the case. Some caution is required however in interpreting these plots, since the scale is such that the correlation for the first few hours cannot be clearly seen, and the white noise boundary is not plotted. A figure shows the variation of correlation with distance between sites, both for the original and synthesized data. There is a satisfactory overlap between the two sets, although the model generally results in slightly higher correlations. In contrast, the degree of agreement between historical and simulated cross-covariance coefficients is very high for the Bath model, as described in section 4.1.2. This superiority could be a consequence of the higher model order chosen for the Bath model.

Forecasting accuracy was calculated for the VAR model of [132], for up to 6 hours ahead, and was compared with persistence forecasts. Plots for several locations show significant improvements, with percentages varying from around 14% at 3 hours ahead to 19% at 6 hours ahead. It was observed that three out of the four sites that gave the best forecasting performance were of lowland terrain, so terrain type clearly has an influence on model performance. The improvements of VAR over univariate results are substantial and consistent across all sites and look-ahead times. A figure for one site shows that the model has preserved the shape of the annual wind speed distribution, although not perfectly - the mode is 2m/s higher in the simulated data, for example.

The ability of information from multiple locations to improve forecasting was also explored by Alexiadis, Dokopoulos and Sahsamanoglou [133]. Limiting themselves to a situation in which there is strong dominance by a prevailing wind direction, the authors construct a 3 layer artificial neural network model which predicts the wind speed at a reference location based on both the location's own past values and values at another location upwind from it. With relatively short prediction times and the sites separated by distances of the order of tens to a few hundred km, the method was able to outperform persistence by 20 – 40%, which is very significant compared to models based on only one location.

In a research project commissioned by generation company E-ON, Allwrite [134] was concerned with producing synthetic wind speed time series for a set of real wind-farm sites, ensuring that all major statistical properties are preserved. These are listed as: (i) cross-correlations that are a negative exponential function of distance; (ii) correct autocorrelation functions and (iii) Weibull marginal distributions. Data was provided by generation company

E.ON and consisted of 10-minute average wind speeds from 7 wind farm sites. Data coverage was rather sparse, resulting in only a 3 month period in which there is synchronized data for 5 of the sites, so the task was to fit a 5-variate time series. Correlograms were plotted, and the CCFs show, as intuition suggests, that the cross-correlations do not have their peaks at zero lag, rather a few hours.

Looking first at the univariate case, models of various kinds are proposed, including GARCH, but the type chosen involved simple AR models with a complex Gaussian random variable $X_t = X_t^{Real} + iX_t^{Imag}$, with the two components independent and of equal variance, and letting the measured wind speed $U_t = |X_t|$. This means that U_t will be Weibull distributed with shape factor 2, but can be converted to any shape factor k_W by raising to the power $2/k_W$. The multivariate case proceeded similarly, with $\underline{X}_t$ and $\underline{Z}_t$ as 5-dimensional complex vectors, and letting $\underline{U}_t = |\underline{X}_t|_c$, where the subscript c means that we take the absolute values component-wise to get the vector of simulated speeds.

Autoregressive models of increasing order were fitted to the multivariate time series, initially assuming Gaussian variables and using the Yule-Walker equations. Noting that most of the reduction in the sum of the residuals' variance is obtained by the first step – i.e. moving from an AR(0) to an AR(1) model, it was decided that AR(1) was sufficient. Model fitting then took place for the complex model using the principle behind the Yule –Walker equations, i.e. matching covariance values with lags 0 and 1, while noting that it is quicker to match the 4th moments.

A similar approach was adopted by Correia and Ferreira de Jesus [135] when developing an algorithm for fitting and simulating VAR(1) models for wind farms, with each series representing the wind speed experienced by individual turbines within the farm. The procedure does not involve transformation of the series to be Gaussian - they are assumed to be Weibull distributed, with different coefficients for each zone. Use is made of the fact that a Weibull distributed random variable X_t with shape and scale parameters k_W and C_W may be expressed as the sum of two independent $N(0, \sigma^2)$ distributed variables $X_{1,t}$ and $X_{2,t}$ according to

$$X_t = (X_{1,t}^2 + X_{2,t}^2)^{1/k_W} \text{ where } \sigma^2 = C^{k_W}/2. \quad (4.2)$$

A comparison of real and simulated data shows that autocorrelation is considerably over-represented for lags greater than 1, attributed to the limitations of a 1st order model. The methodology does not seem to allow easy adjustment of the algorithm in order to increase the model order.

Klöckl [136] developed a VAR model for hourly averaged wind speeds, arbitrarily using data from Arizona. His methodology was to work explicitly with cumulative distributions to transform each wind speed time series to being exactly Gaussian, fit a VAR model using standard methodologies, and to reverse the transformations for series simulation. Additionally, he assumed that each hour of the day had different wind speed distributions – a decision probably taken since he was modelling both wind speed and solar irradiance for system studies. This is the methodology discussed in Chapter 3, equation 3.21, and Klöckl does not discuss a possible need to use the more demanding transformation of equation 3.22. An advantage of this methodology over the simpler power transformation is that it avoids the need, in the latter case, of dealing with negative synthetic wind speeds which will inevitably be generated. A rather coarse way of dealing with such values is to set them to zero as a last step.

The methodologies above do not question the implicit assumption that inter-dependencies between series can be considered fixed and linear (except perhaps for the briefly described neural network model). One piece of research which does not rely on such assumptions was conducted by Grothe and Schnieders [137], concerned with the optimal allocation of wind farms across Germany. A model was developed that can, for a given allocation of capacity to a set of locations, assess the lower quantiles of the distribution of the overall produced wind power. Algorithms were developed to maximize these quantiles by re-destribution of capacity, subject to certain constraints, to obtain optimal allocation plans for wind energy production. The motivation is that if we are interested in providing a stable baseload, the lower quantiles are more important than variance. While optimization of the variance only requires estimates of the marginal variances and covariances, whenever multivariate data is not joint-normally distributed, as appears to be the case for wind speeds, the quantiles of sums of margins may not be calculated from sums of variances and covariances. The assessment of quantiles for the joint distribution the aggregated power is not trivial, requiring modelling of the marginal distributions and their entire dependence structure. This may be achieved through copula functions, described in Chapter 3, section 3.5.

Historical datasets were provided by the German weather service, consisting of daily mean wind speeds measured at 40 locations and hourly speeds measured at 39 locations. Both onshore and offshore weather stations were included, for the period 2005 to 2010. As the datasets were described as highly skewed this was removed by applying the Box-Cox transformation, with the transformation parameter estimated by maximum likelihood. Statistical tests (Ljung-Box and Engle's ARCH test) strongly indicated the presence of auto

regressive structure and heteroskedasticity (changing variance) in the data. Univariate models were developed to clean the time series from these 'effects': seasonal ARMA models, with seasonal functions S_t for their means, and seasonal volatilities σ_t to account for heteroskedasticity. The resulting standardised residuals ξ_t for each of the time series were described as passing Engle's ARCH test, thus showing "no significant heteroskedasticity" [137]. This means that the null hypothesis - that if regressing squared residuals on their recent past values, the coefficients will all be zero – cannot be regected to some (unspecified) high level of confidence.

This is at odds with the conclusion of other authors that conditional heteroskedasticity exists in wind time series that cannot be deterministically removed. However, the difference can probably be explained by the fact that in [137], values were daily rather than hourly averaged.

The authors model the nonlinear dependence structure of wind speeds at different locations through the dependence of their concurrent model residuals. To justify this approach, they had to ensure that such residual dependencies are consistent in time and can capture the complete dependence of the series. For the first point, they looked at the pairwise cross-correlations of the empirical residuals for lags -15 to 15. The approach is justified if there is no statistically significant cross-correlation for non-zero lags, which turned out to be the case for the daily series, but not the hourly ones. In the latter situation, the approach is not adequate since lagged residuals contain information which is not captured by the dependence of simultaneous residuals. The authors looked at the temporal evolution of the rank correlation coefficients between pairs of wind speed residuals within a rolling window. They found that although the correlations vary in time, they stay within fairly narrow limits and can thus be reasonably modelled as constant in time.

The authors decided upon the types of parametric copula functions that best approximates the dependencies between pairs of residual series with the use of plotted dependency functions. Example plots show clearly that the dependencies are non-Gaussian and heterogeneous, i.e. the type varies across pairs of series. The heterogeneous structure together with the high dimensionality of their problem complicated the process of finding of adequate copula functions, as most copula functions assume homogeneous dependency structures across dimensions. Therefore, the authors chose to use multivariate pair-copula constructions based on a hierarchical tree of two-dimensional copulas. Since the simultaneous estimation of the marginal models and the copula structure by maximum likelihood was

computationally very complex for 40 dimensions, the authors first estimated the models for the univariate time series and then used the corresponding residuals to compute the copula structure.

When simulating power outputs for the location optimisation process, synthetic wind speed data were transformed into powers through up-scaling to hub height using the log law, before application of a simple representative function for a turbine power curve.

4.3. Markov Chains

Many authors have chosen to model wind speeds as discrete Markov processes, and produce synthetic wind speed time series through Markov chain Monte-Carlo simulation. Some examples are presented here, all of which are univariate.

Brayshaw et al. [83] modelled wind speeds at two locations in GB, based on hourly averaged UK Met Office data, and created separate models for 3 states of the NAO (low, medium and high), as discussed in Chapter 2. These were Markov chains with a total of 31 states, obtained from dividing the observed range of wind speeds into 1.5m/s wide bins. Precise wind speeds were generated using a second random number, based on the assumption that the probability density within each state is constant. An auto-correlogram up to a lag of 5 days shows that real and synthetic data have correlation functions that are similar for the 1st 24 hours, but that the model underestimates persistence significantly after that lag. The authors state that this could be overcome by fitting a higher order model, but that the fitting of such models is 'extremely challenging'. A constant up-scaling factor was used to translate wind speeds to hub height, and a power curve used to convert them into single turbine power outputs. The variance of power output for synthetic and real wind speed data were compared for averaging periods ranging from hourly to monthly, and found to be very similar. The similarity is interpreted by Brayshaw et al. as suggesting that multi-day persistent low-wind or high-wind events are being represented in the model to some extent, but the model probably fails to capture the most extreme cases of either.

Brokish and Kirtley [138] seek to determine when Markov chains are appropriate for modelling wind, and demonstrate the shortcomings of inappropriately applied Markov models. They summarize 7 Markov-chain models for wind speed, ranging from the simplest one which is 1st order with 3 states, to the most sophisticated which is 3rd order with 35 states. The averaging periods for the wind speed series used varies from eight hours to less than 0.3 seconds. A common thread of discussion in these studies is the fact that ACFs are not reproduced accurately, and that the problem is much worse for models using short averaging periods. The reason might be that a fixed lag of e.g. 36 hours involves a much higher number of steps for 5 minute averaged speeds than hourly averaged ones.

A metric was introduced to compare the quality of fit between synthetic and real auto-correlograms – the sum of r.m.s errors up to a lag of 12 hours. The metric showed that increasing the model order improves the match, but to a limited extent – the improvement being most pronounced for sub-hourly averaging periods. Examining the number of states, the

authors found that, as expected, more states generally resulted in less error – although the sensitivity is not high, and for 80 minute averaged data an 8 state model performed better than an otherwise equivalent 16 state model. The authors ran simulations to calculate storage requirements for a wind driven micro-grid with real wind speed data and synthetic data from several models, finding that the underestimation of persistence lead to underestimation of storage requirements, in some cases gross underestimation, although the effect is worst for sub-hourly averaged data.

In contrast, it was found by Hocaoglu, Gerek and Kurban [139], modelling data recorded in Turkey, that the number of states does have a significant effect - although they examined distribution statistics rather than persistence. Two 1^{st} order Markov chain models were constructed, one with 13 states (1m/s bins) and the other with 26 states (0.5m/s bins). Examining mean, median and standard deviation from the real data and simulations of both models it was found that the percentage errors for two models were -10% and +2.6%, respectively, for the mean; -12% and +3.6% for the median and -9.5% and -4.3% for the standard deviation.

Shamshad et al. [140] fitted 1^{st} and 2^{nd} order 12 state Markov chain models to hourly averaged wind speed data recorded at two locations in Malaysia. The mean, standard deviation, minimum and maximum values and the percentiles of the synthesized values are presented together with the observed ones, showing good agreement and, in this case, no significant improvement for the second order model compared to the first order one. The Weibull distribution parameters were also computed for the observed and generated data. The authors claim that both models have, in general, preserved Weibull parameters - although in fact the shape parameter is significantly underestimated, particularly for one of the sites.

Examination of the auto-correlogram, for both models and real data, shows once again that persistence is under-represented, the main divergence between real and synthetic data happening at about 6 hours lag. The 2^{nd} order model performs better up to a lag of about 24 hours. Spectral analysis was carried out, leading to the plotting of a kind of periodogram – with hours/cycle rather than frequency along the horizontal axis, from 1 hour/cycle to 700/cycle. The general shape of these three curves is similar, the second order model performing slightly better. Both Markov models however have constant spectral intensity from about 48 hours/ cycle, whereas the observed data has a gentle positive gradient throughout the range, indicating long memory.

A univariate semi-Markov model of wind speed was developed by Negra et al. [141]. More precisely, they developed a model of a type described as 'birth and death'. In common with discrete Markov chain models, wind speed is modelled as being in one of a number of states, i.e. lies within a certain range. Further, during a transition between states, the wind speed may only move up one level (birth) or down one level (death). However, as opposed to discrete Markov chain models, the time spent within a state is continuous, and follows an exponential distribution (where probability density is a negative exponential function). The probability of a transition from a given wind speed state to another state is directly proportional to the long-term average probability of existence of the new state. In order to preserve seasonal characteristics of the measurements, separate probability tables were defined for each month of the year.

In order to verify model performance, histograms of both real and synthetic datasets were plotted, and found to be in very good agreement. Correlograms with the ACF's for several real years were plotted, along with synthetic datasets, up to 100 hours. These demonstrate that a second- or third-order correlation must be included in the model, i.e. the movement of wind speed at hour t depends on the values of wind speed at hours $t-1$, $t-2$ (2^{nd} order) and $t-3$ (3^{rd} order), and conditional transition rates must be defined. The best choice of order is said to depend on the amount of input data, with both solutions able to provide satisfactory results. However, it is clear that second-order correlation produces a slightly under-correlated series, while it is over-correlated for third-order models, although both offer reasonably good fits.

4.4. Hierarchical Models

4.4.1. Markov-Switching Models

Ailiot, Monbet and Prevosto [142] are concerned with modelling the wind velocity field in a region of the north east Atlantic, as measured at a set of 35 points, and develop a hidden Markov-switching autoregressive model. They argue that linear autoregressive models describing the time series of wind fields cannot reproduce some features of the wind fields, in particular the motions of meteorological structures such as cyclones. The authors state that AR models can successfully describe the motions of objects that translate at a constant speed, but motions of meteorological structures depend on the position of principal air masses and evolve in time. They alternatively propose an model in which these motions are introduced as a hidden Markov chain. Conditional to this hidden process, the evolution of the wind field is modelled using autoregressive models with time-varying coefficients. No clear evidence is provided of the extent to which this model structure improves wind velocity modelling as opposed to a static model.

Pinson and Madsen [143] examine the output of large Danish offshore wind farms with 10-minute average resolution. They found that with a large amount of capacity concentrated in one area there is insufficient spatial smoothing to eliminate large fluctuations. Quoting several authors, they state that there exist sudden changes in the fluctuations' characteristics – essentially their frequency and magnitude, at time scales of a few hours. These changes, they claim, cannot yet be explained in terms of the evolution of an explanatory variable and are best represented by the hidden (Markov) switching of states determining AR model coefficients. The authors refer to studies which used a sliding window fast Fourier transform approach to establish that the spectral characteristics of offshore wind speed exhibit frequent and abrupt changes, which a hidden Markov Model can capture, but also smooth variations at the time-scales of months and seasons. The latter, the authors believe, must be reflected in the model and can be achieved through an adaptive parameter estimation method, involving exponential forgetting of past values.

The Markov-switching models developed by them were assessed for their out-of-sample forecast accuracy, finding a reduction in normalised RMS error of 7.5% over persistence and 2.6% over simple AR models for 1-step-ahead prediction. Although improvements are fairly modest, they are seen as satisfactory given that persistence is difficult to outperform for 1-step-ahead prediction. Pinson and Madsen [143] also demonstrated that their models are also useful for producing predictive densities, consisting of finite mixtures of

conditional densities in each regime. It is admitted however that parameter uncertainty has not been considered and may be an issue as the adaptive estimation framework means that the quality of estimation may vary with time. It is hoped that in the future, the regime sequences may be compared with the time series of other meteorological variables, in order to establish an explanation for the changes in terms of those variables. One possibility is that the hidden states might correlate with the circulation types discussed in Chapter 2, although one immediate obstacle is that the modelling methodology of this paper found that only 2 or three hidden states are optimal, which is smaller than the number of states even in Brawshaw and Masato's reduced scheme [77]. It seems likely that differences between regimes would be reduced if the data were grouped into hourly averages.

4.4.2 Use of Wavelet Analysis

As established in Chapter 3, such phenomena would be better analysed in a wavelet transform context. That is the approach taken by the LETS Project: Locally stationary Energy Time Series, a fairly small consortium mainly involving statisticians from Bristol and Lancaster Universities. At the time of writing, no literature has been published by them of direct relevance to this research.

The use of powerful wavelet techniques such as the DWT and MODWT in the analysis of multivariate meteorological time series is described by Whitcher, Guttorp and Percival [144]. Such methods would probably be very helpful to the process of finding explanatory variables for the non-stationarities described above.

4.5. Producing Wind Power Outputs from Wind Speed Time Series

Gibescu, Ummels and Kling [146] were concerned with generating synthetic time series of wind speeds at a set of likely future wind farm locations in the Netherlands. The intention was to convert them to power outputs, concurrent with historical wind time series for a set of present wind farm locations. The task is therefore one of spatial interpolation modelling. The historical time series were 10 minute average wind speeds over 1 year at 12 locations - 9 onshore, 3 coastal and 6 offshore. Examination of these datasets showed that their (long term) variances are a linear function of their (long term) means – a problem rectified through the replacement of wind speeds with their logs, so that after the removal of means, the series could be modelled as zero-mean stochastic processes with equal variance. The means to be removed were assumed to be functions of time of day, but not time of year, and it was found that offshore sites have a strong unimodal pattern, with hardly any pattern at onshore sites. The mean pattern for the sites to be synthesised were calculated as linear interpolations of the patterns at the known sites.

For a given hour t the set of all wind speeds, i.e. the known historical ones and those to be simulated, are treated as a multivariate normal – despite acknowledgement that their marginal distributions are not truly Gaussian. The cross-correlations for historical sites were easily calculated, and the best fitting $exp(-d_p)$ curve for them (as discussed in Chapter 2) enabled extraction of cross-correlations between pairs of simulated sites, along with historical-simulated pairs. The simulated values are then taken to be the expectation values of the unknown elements of the multivariate distribution, conditioned on the known values for that hour. Power outputs for the wind farms were calculated using a smoothed power curve, taken from the website of wind turbine manufacturer *Enercon* (reference within [146]), which is described as accounting for spatial smoothing and wake effects. The power curve is quite different from that of a single turbine, rather surprisingly different given the small spatial separation of turbines in a single wind farm.

The distribution of errors for log wind speeds generated using this method were found to be Gaussian, making the production of confidence intervals for the power outputs relatively straightforward. The standard deviation of the wind power production conditional on the observed wind speeds, i.e. the average final uncertainty in the method, came out as about 20% of the average amount simulated – a rather large figure resulting from the highly nonlinear power curve.

Nørgaard and Holttinen [147] were concerned with the common situation whereby information about the instantaneous wind resource for an area is available in terms of a time series of wind speed, valid only for one specific site, but representative of the whole area. This is the case for the model developed in the current research, where a single wind speed represents a GB zone – although this may in fact be a set of 3 or 4 proportional wind speeds if we up-scale differently for different location types, as was the case in the more advanced Bath methodology. The paper presents the outline of an algorithm for generating a time series of aggregate power output from multiple similar turbines from this single-location wind speed series, based upon consideration of smoothing effects in both time and space.

The algorithm involves the creation of a multiple turbine power curve and requires definition of a characteristic length for the area, the average wind speed, and a rough figure for the turbulence intensity. The authors claim that the methodology can be applied to areas ranging from only a few km in extent (i.e. individual wind farms) to several hundred (i.e. a region) although it seems some adjustment may be necessary for hourly averaged series and individual wind farms. The methodology was developed as part of a large international academic project funded by the EU. The project was called WILMAR (Wind Power Integration in Liberalised Electricity Markets) and was aimed at improving understanding of issues related to the integration of a significant wind capacity into integrated European electricity markets. The algorithm, as applied to data in the current research is described in Chapter 8.

Complimenting this work is an excellent study by Holttinen [55] on the expected statistical properties of the power output time series for large-scale, geographically spread wind capacity. Her conclusions lead to guidelines, rather than a procedure, for the transforming of a set of single-location wind speed series to large-scale power outputs, avoiding inaccurate up-scaling of variability. The study made use of the extensive available hourly time series for Denmark, and more limited, but wider area data from other Scandinavian countries. Holttinen reports that:

- For a single turbine the standard deviation is somewhat larger than the mean, about 30% or even 40% of capacity, but for a European country the standard deviation should be about 20% of capacity;
- Relative to mean production, this standard deviation should be 0·5–0·8 for a circle of radius 200 km, 0·4–0·6 for radius 1000 km, and saturate at about 0·3 when the radius gets larger than 2000 km – which is beyond the dimensions of GB;

- Hourly variations should be within ±20% of capacity, or even less if the area is larger than the size of Denmark – as is the case for GB;
- The standard deviation of the series of hourly changes, for a country, should be less than 3% of capacity;
- The maximum hourly production should be less than 100% of capacity: 85% – 95%, depending on how large the area is. Examples: Denmark 93%, Finland 91%, Norway 93%, Sweden 95%, entire Nordic area 87%; and
- The duration of calms across a country, i.e. periods where production is below 1%, should be non-existent or limited to about 5%, for a country the size of Denmark.

4.6. Direct Modelling of Aggregated Wind Power Outputs

There are very few published studies which fit models directly to the power output of large aggregates of wind generators, rather than single locations. One exception is Sturt and Strbac [56] who fitted time-series models directly to the hourly averaged, aggregated wind power outputs of nation-sized wind fleets - New Zealand, Denmark and Germany. The usual method of fitting a multivariate model such as VAR, the authors state, can run into 'calibration difficulties' due to the very large number of parameters that are required if many sites are to be represented.

Their approach is to use a Gaussian, low order autoregressive process as an underlying driver, then transform it by adding a periodic diurnal term, and then a non-linear function designed to give the process the exact required asymptotic distribution. This function is sigmoid shaped and relies upon a false assumption that wind-farm power curves are monotonically increasing. Annual variations are accounted for by dividing the year into a number of seasons and fitting separate models to each one. The underlying process is filtered Gaussian noise, scaled so as to normalise the process' distribution to $N(0,1)$ when the autoregressive coefficients have been manipulated so as to reproduce the desired transitional statistics.

The authors state that while the beta distribution has been suggested by several sources for describing aggregated wind output, they prefer to estimate the distribution non-parametrically using Parzen windowing. An algorithm is presented which allows determination of the nonlinear functions and vectors of means required to satisfy the required asymptotic long-term statistics of the power output. The transformation function is represented as a piecewise linear approximation with a few hundred points.

The model structure was fitted to historic wind power data from New Zealand, Denmark and Germany. These cases provided contrasting challenges for the model, with a small number of highly dispersed sites in New Zealand, a large number of sites in a much smaller region in Denmark and a large installation in Germany. It was found that while a first-order model is adequate for the New Zealand data, second-order ones are necessary in the case of Denmark and Germany. Autoregressive parameters were fitted using standard methods such as the Yule-Walker equations, but another important aspect of model calibration was visual comparison of simulated and historical time series plots, leading to some manual parameter adjustment. The fact that some manual adjustment was required

demonstrates that the best fit in the underlying Gaussian domain does not necessarily lead to the best fit in the power domain, because of their nonlinear relationship.

Four crucial plots are provided for each country, for both historical and simulated data: the relative occurrences of (binned) power outputs, the mean absolute changes for time horizons up to 24 hours, the distributions of changes on both hourly and 4-hour horizons. In the synthetic case, the plots show the mean curves from 1000 simulated years, along with 97.5^{th} percentile (upper bound) and 2.5^{th} percentile (lower bound) curves for those years. For historical data, curves for individual years are shown (2 – 4 years), and it is seen that the historical data fall mostly within the bounds. This is claimed as a probable validation by the authors, although given the significant differences between the very small number of historical years, it seems likely that a sample of 1000 historical years would contain many more extreme years that were mostly not contained within the bounds – i.e. inter-annual variations are surely rather under-represented by the models, but it is impossible to know to what extent.

A roughly log-linear relationship is observed for the distribution of sudden changes for Germany and Denmark, indicative of a Laplace distribution as, they state, has been noted by several authors for regions of various sizes. The New Zealand data however seems to be sub-Laplace, which should be expected for very large regions due to the Central Limit Theorem.

It may be important to reproduce the relationship between volatility and power level, for system simulation. The model does this quite well, although it under-predicts the volatility at very high power outputs due to its inability to represent turbine cut-out. The authors also tested the model's ability to reproduce the incidence of calm periods of various lengths, defined as the length of time spent with a generation level below 5% of total capacity, for Denmark and Germany, and 10% capacity for New Zealand (due to the superior wind resource there). In all regions, the model under-predicts the frequency of calms of 2 hours or *below*, ranging from a 15% under-prediction for New Zealand, to a 30% under-prediction for Denmark. There is generally good agreement in the distribution of calms of between 2 hours and 7 days, although in the case of Denmark the frequency of very long calms is under-predicted by the model. The extent of agreement is encouraging, given that the model's driving process has such a low autoregressive order, and it is interesting that a lack of long memory representation has limited impact.

Sturt and Strbac applied the same methodology to develop four seasonal models representing the aggregate output of a possible British wind fleet circa 2030, presented in

[148]. Since the model is of a hypothetical distribution of wind generation capacity, and the real GB wind output series available to date are from capacity concentrated almost entirely in southern Scotland, the model was calibrated to series derived from Met Office data. Since no offshore data was available, the offshore capacities were initially mapped to nearby onshore regions. Regional weightings were taken from the core scenario for 2030 presented in an influential report by Pöyry Energy Consulting on the impact of wind intermittency [149]. (They are a private consultancy firm often hired by the UK Government to help answer challenging technical questions relating to energy policy and markets).

Choosing AR(2) for the underlying process, exact fits were achieved between historic and simulated values for the histogram of wind power outputs and the variation of means by season and time of day – as a natural consequence of the methodology. The model also provides a good fit to all the historic power output change distributions, including the most extreme events occurring once per year or less. However, analysis of annually averaged capacity factors shows that the historic dataset exhibits significantly more inter-annual variation than the simulation. The range of annual capacity factors across the six years of historic data was 24.3%–30.3%, but analysis of many six-year-long simulations revealed that in only 3% of cases did the range of annual capacity factors match or exceed this range.

Examining the available data for metered wind turbines in southern Scotland indicated clearly to the authors that using a fixed speed-up ratio to convert wind speeds from anemometer-height speeds to turbine hub height leads to an excessive diurnal variation in the power output. This is due to the dynamic nature of stability conditions in the boundary layer, as discussed in Chapter 2, essentially differing amounts of momentum transfer due to convective vertical mixing. The authors state that offshore wind turbines do not display diurnal variation in mean power output, since the sea surface is not heated by the sun in the same way as the ground. Therefore, given that roughly half of the capacity in the 2030 scenario is offshore, along with the overestimation of onshore diurnal variation, the additive terms which gave rise to the variation were reduced by 75%.

Another important difference between offshore and onshore wind power is the higher average capacity factor of turbines at offshore sites. To account for this difference, Sturt and Strbac again refer to the report [149] by Pöyry Energy Consulting. The report provides an estimate for the aggregate wind power output distribution of future GB fleet of wind farms assumed (by both Pöyry and Sturt and Strbac), in the form of duration curves. In order to

adjust their model to be consistent with the curves from [149], Sturt and Strbac suitably "stretched" their (seasonal) nonlinear transformation functions.

They did not adjust the autoregressive parameter, despite the observation that the effect of increasing region size is generally to reduce the short-term volatility in the Gaussian domain, and they may therefore be somewhat overestimating the short-term volatility of a true aggregate of onshore and offshore output. Interestingly, there are also reasons for supposing that the reduction in volatility is limited - wind speeds tend to be more coherent offshore, and large groups of turbines will be located close to each other, as is mentioned in the previously discussed National Grid report [57]. Unfortunately an analysis of the goodness of fit of the adjusted model, particularly the transitional statistics, was not possible due to the lack of high-quality offshore wind time series.

Prior to these research projects, Professor Strbac worked with the consultancy company Ilex to quantify the additional system costs likely to be incurred if the volume of renewable in GB were to increase to 20% or 30% of demand by 2020. Their analysis [150] suggests that using wind speeds and power curves overestimates generation and underestimates intermittency – so the study only uses half-hour metered generation data from UK wind farms. One of the report's conclusions is that there is as much variation in output within regions as there is across them – a result that seems somewhat at odds with other research described here, perhaps as the data was obtained from an insufficient amount of wind generation capacity.

4.7. Modelling Long Memory and Conditional Heteroskedasticity in Wind Speed Time Series

It was shown in Chapter 2 that wind speed time series display long memory, as found by Haslett and Raftery [88] when modelling the wind resource in the Republic of Ireland. Chapter 2 also presented an argument by Bouette et al. [90] that wind speeds should rather be modelled as seasonal long memory, or Gagenbauer processes. Chapter 3 described how both regular and seasonal memory can be represented through the use of fractional differencing operators, and this section will discuss in detail how wind speeds were modelled in these papers – probably the most influential in setting the course of this research. It was also stated in Chapter 2 that wind speed series display stochastic conditional heteroskedasticity, as described by Tol [91]. This section presents another paper [152] of considerable significance to this research, which reports on the development of a regression based model of wind speeds that incorporates both long memory and conditional heteroskedasticity.

4.7.1. Examples of ARFIMA and GARMA Models

As previously stated, Haslett and Raftery [88] were concerned with establishing the wind power resource at sites in the Republic of Ireland for which only short time series of daily averaged wind speeds were available, based on similar but long term records at 12 sites across the country. They realised that a long memory model was necessary due to the excessive variance of means for extended periods such as a month. This appears to be in contradiction to the findings of Brayshaw et al. [83], but this is not the case since he was comparing synthetic and historical variances for the same NAO states, while Haslett and Raftery worked with variances for periods that represent a wide range of NAOI values. This certainly reinforces the view that long memory and hidden state switching are two perspectives on the same phenomenon

The approach taken by Haslett and Raftery was to first de-trend the data by calculating and then removing a function representing the annual seasonality, consisting of a superposition of trigonometric functions. Obviously, the diurnal seasonality is not observed due to the daily averaging period. Once detrended, the data is modelled as a fractionally integrated ARMA process, following equation 3.16 in Chapter 3. A single univariate model is assumed to apply to all sites – i.e. the same model order and parameters are taken as the compromised best fit for each location, in order to reduce the number of parameters to be estimated. However, to allow for spatial correlation the white noise innovations are assumed

to be generated by a multivariate normal distribution. Here they are making the same assumption as Grothe and Schnieders [137], that the series' spatial dependencies may be fully captured by the dependencies of their concurrent model residuals – validated as true for daily, but not hourly averaged data. The correlation between the innovations at any pair of sites is given by a negative exponential function of the distance between them, with coefficients to be established empirically.

Model identification and preliminary estimation proceeded in the following way (with a change from the original nomenclature):

1. Preliminary estimates of coefficients were fitted to describe the relationship between correlation and distance, allowing construction of the covariance matrix Σ_{HR} for the white noise series.
2. Establish the matrix $\boldsymbol{C_\Sigma}$ satisfying $\boldsymbol{C_\Sigma \Sigma_{HR} C_\Sigma^T C = I}$. After pre-multiplication by $\boldsymbol{C_\Sigma}$, the series are assumed to be spatially independent. This makes the problem to be solved univariate.
3. Apply an AR(9) filter to the datasets, to remove short memory dependence and leave residuals with persistence, i.e. ARIMA$(0, d, 0)$ processes.
4. Obtain an estimate for d from the slope of a plot of log variances for sample means of the filtered series vs. sample size.
5. Return to the series before AR(9) filtering and fractionally difference it, using the estimated value of d and binomial expansion.
6. Identify a common ARMA(p,q) model for the differenced series. An AR(2) model was identified.
7. Carry out maximum likelihood estimation of all the coefficients. This is a computationally expensive process, particularly back in 1989, so to simplify use was made of the fact that conditional means and variances may, according to the authors, be found to a good approximation using only the partial autocorrelations for the ARFIMA$(0, d, 0)$ process.

The process resulted in an estimate of d = 0.328, implying a strong presence of long memory. Values found for the 1st and 2nd autoregressive coefficients were 0.01 and 0.063, respectively – very different to persistence and the short memory models developed at Bath. Rather unexpectedly, the model is close to ARFIMA$(0, d, 0)$, with the long memory parameter accounting for most of the correlations for short and long lags.

The arguments made by Bouette et al. [90] in favour of modelling daily averaged wind speeds as a Gagenbauer process were discussed in detail in Chapter 2 (section 2.5.1), and the associated GARMA model structure and process simulation were described in Chapter 3 (section 3.2.1). When estimating the model parameters for Roche's Point, one of the stations from Haslett and Raftery's set of 12, an obvious pole at the annual frequency confirmed that the seasonality parameter $\omega_G = 2\pi/356$. The parameter δ dictates how steeply the spectral intensity function rises close to the unbounded frequency, and as a result one can construct an estimator for it based upon the periodogram in this region - doing so gave Bouette et al. δ = 0.15. The authors also made use of an approximate maximum likelihood estimator developed by Whittle, presented in Beran [86] and will be discussed in Chapter 7. This estimator is based on the discrete periodogram for the process, along with the theoretical spectral density at the same set of frequencies, given a full set of model parameters. This method gave a value of δ = 0.18, so there was fairly good agreement between estimates.

Additionally, for $v = 1$, which is nearly the case, the differencing operator for a Gagenbauer process reduces to that for a simple long memory process with $d = 2\delta$, so Bouette et al.'s results might be in good agreement with that of Haslett and Raftery. It seems more likely however that the latter involved an over-estimation of long memory, arising from a lack of computational power.

Although Whittle's quasi-log-likelihood estimator is based on the full set of parameters for the GARMA model, the authors appear to have only used it to obtain a value for δ, stating that it is very difficult to estimate the ARMA part of the model since "one needs to invert the Gegenbauer part of the process" in order to do so. The paper does not discuss the best choice of autoregressive and moving average model order, nor disclose any parameter values. As stated in Chapter 2, the authors' suggestion that hourly averaged series should be modelled as 2-factor Gagenbauer processes is rejected, since the interection between the two seasonalities would not be captured.

4.7.2. Inclusion of Conditional Heteroskedasticity

The last paper to be discussed is this review is by Caporin and Pres [152]. These authors consider the wind resource in terms of risk exposure, primarily for generation companies, and wish to develop a univariate model which allows simulation and probability density forecasting for the wind speed at single locations. They consider a set of models that have appeared only rarely in the time series literature, generally in quite different contexts:

ARFIMA-FIGARCH (proposed by Beine and Laurent in 2003), and two competing approaches that introduce serial correlation in Gamma densities.

In contrast to other researchers, Caporin and Pres do not base the models on data transformations such as the Box-Cox, but on actual original data, for which they postulate a specific stochastic structure whose components have a direct interpretation. In all their model choices they assume that the periodic components are purely deterministic, and may be filtered out a priori. Doing so allows better modelling the underlying stochastic process, they believe. They note, however, that they could have followed the alternative approach of assuming a stochastic nature for the periodic components, allowing specifications such as GARMA, SARMA or SARIMA models. The comparison of their modelling approach with those considering stochastic periodic components is left to future researches, which undoubtedly includes the current project.

Caporin and Pres propose two alternative specifications for the deterministic seasonality to be removed. If the original seasonal series is X_t, and the deseasonalised series is X_t^{Stat}, the first approach is to assume

$$X_t = \exp(m(\mathrm{t}))X_t^{Stat}, \quad \text{so that} \quad \ln(X_t^{Stat}) = \ln(X_t) - m(t), \tag{4.2}$$

where $s(t)$ consists of a polynomial and sinusoidal functions. The 2nd approach is to assume that

$$X_t = \exp(m(t) + s(t)X_t^{Stat}), \quad \text{so that} \quad X_t^{Stat} = \big(\ln(X_t) - m(t)\big)/s(t), \tag{4 .3}$$

where $s(t)$ has a similar structure to $m(t)$. The 1st specification requires a stochastic model for positively defined random variables only, whereas the 2nd requires random variables with support over the real line. The ARFIMA-FIGARCH model assumes a normal distribution for innovations, corresponding to a lognormal distribution of wind speeds, and as such requires the 2nd specification of seasonality.

It was observed in the preliminary data analysis conducted by Caporin and Pres that some long-memory may be present in both the mean and variance of the seasonally adjusted series – albeit with weak evidence in both cases. An ARFIMA–FIGARCH process allows for both. As the name suggests, the ARFIMA-FIGARCH specification assumes that the process mean follows the ARFIMA definition, i.e. equation 3.16 from Chapter 3 with $0 < d_{mean} \leq 0.5$, while the conditional variance is similar to GARCH, equation 3.35, except with the addition of long memory. More precisely, the behaviour of the conditional variance h_t is governed by

$$h_t = \alpha_0 + \beta(B)h_t + [1 - \beta(B) - \Psi_F(B)(1 - B)^{d_{var}}]z_t^2. \tag{4.4}$$

The model was fitted to three historical daily-averaged wind speed intensity series for meteorological stations in Poland, from 1986-2008. Four interesting plots are included in the paper, derived from one of the 3 series: the entire 23 year series plotted, a correlogram showing the ACF up to a lag of 1000 days, the periodogram and kernel density estimate. The ACF initially decays very quickly (on the scale of the plot) before settling into a sinusoid, with an annual period, which appears to decay slowly with an average level that remains above zero even after nearly 3 years. The periodogram has two very strong peaks, one at zero frequency and another, significantly larger at the annual frequency. The density estimate, they state, appears to be consistent with either Gamma, Weibull or log-normal distributions. Results for the other two locations are said to be similar.

The authors suggest the use of MLE approaches or Quasi Maximum Likelihood Estimation (QMLE) methods to estimate the parameters of the model. The estimation of the entire model could be performed in stages, beginning with $m(t)$ then $s(t)$ by OLS, before estimating the ARFIMA–FIGARCH parameters using a normal likelihood function. Given the importance of good starting values when estimating long memory parameters, they suggest initialising the estimation using preliminary coefficients obtained by applying the Geweke and Porter-Husak estimator, or Whittle estimators – described in the next Chapter, to X_t^{Stat} and z_t^2. Here the innovation series is obtained by fitting only the ARFIMA part of the model and is used only to recover the preliminary estimate of the FIGARCH long-memory coefficient. Alternatively, several starting values for the memory coefficients could be used to avoid the convergence to local optima.

The model fitting process gave long memory coefficients for the mean process ranging from 0.17 to 0.2, with only an MA(1) term needed to capture short-memory behaviours. In the variances, it was found that GARCH and FIGARCH specifications are very close, with a clear preference for short-memory structures for one location, long-memory for another while the 3rd was ambiguous.

When performing simulation of wind speeds under the model, one can begin by generating standard normal deviates for the normalised innovations $\xi_t = h_t^{-1/2} z_t$, or by resampling them from the in-sample residuals, which may be more accurate if they are not exactly normal in reality.

Given that the main purpose of the study was to forecast wind speed, the authors chose to compare the models on the basis of their one-step-ahead point forecasts and density

forecasts, as well as in their ability to simulate wind speed sequences. Compared in theses ways, the authors found that ARFIMA based specifications provided better results than the alternative models based on the Gamma distribution. There was no clear favourite between GARCH and FIGARCH specifications. The power law was used to convert speeds to hub height, and a single-turbine power curve used to convert to normalised power outputs.

4.8. Summary and Conclusions

This section summarises the most relevant information gained by examining the considerable variety of statistical wind modelling literature.

Most models assume that wind has deterministic annual periodicity in mean and variance that can be removed. Hill et al. [132] take this approach and believe this renders distributions close enough to Gaussian. Deterministic diurnal periodicity may also be assumed, sometimes specific to the season. Such patterns are not found offshore.

Some models, such as the Bath Wind Model, assume that differences in mean and variance within a season are small enough to be ignored. The Bath wind model methodology also assumes that purely winter distributions are close to Gaussian - probably why it tends to over-estimate capacity factors. Several authors apply the Box-Cox transformation, with the transformation parameter either estimated through maximum likelihood or making use of the fact that the Weibull distribution with $k_W = 3.6$ is almost Gaussian. In contrast, Caporin and Pres [152] model wind as lognormal.

Many authors found that the persistence model is hard to beat for hourly averaged series, and that they are close to requiring 1st order differencing for stationarity. There is considerable disagreement within the literature about the best choice of order for ARMA models, ranging from ARMA(3,1) to AR(1). An appealingly simple approach adopted by some authors is to choose an order between those given by the BIC and AIC. Several Matlab toolboxes have functions specifically for the fitting and simulation of time series models, but they are not general enough for the type of model developed here. Bayesian and frequentist approaches were found to yield effectively identical values. Several authors consider that OLS is a sufficiently accurate methodology, with no need for maximum likelihood estimators.

Although checking model residuals for independence is essential, some caution is required since it was found in [127] that the residuals of a simple model displayed no evidence of missing seasonality, despite much evidence to the contrary. Sturt and Strbac [56] found much value in visually examining simulated series, with consequent ‘manual’ adjustment of parameters.

Hill et al. [132] show that the forecasting improvements due to using a multivariate model are substantial. At the other extreme, Haslett & Raftery [88] fitted the same model

parameters to each of 12 zones, after rendering the process univariate through matrix pre-multiplication – this was necessitated by computational constraints a few decades ago. Several authors assume that, following power transformation, multiple wind speed series from a MVN, with cross correlations following a exp(-d_p) pattern. One example was found [137] which does not assume MVN, rather uses a hierarchical tree of heterogeneous copula functions, following Box-Cox transformations.

With regard to discrete Markov-chain and semi-Markov models, it was found that they can only recreate the ACF of the relevant wind series up to a point, then under-represent persistence, partially because they don't allow for long memory - although this is not acknowledged. Higher order Markov chain models generally perform better, but can over-represent the ACF for all examined lags. Apart from this shortcoming, it was found by Brayshaw [83] that a simple 1st order discrete Markov-chain model with 31 states reproduced series quite well. Brokish and Kirtley [138] surveyed many models, and found that while increasing order and number of states leads to better models, the amount of improvement is somewhat limited for hourly averaged winds, with other authors agreeing. Hocaoglu et al. [139] found that number of states is more significant if comparing distribution statistics rather than persisitence. A birth-and-death model [141] seems to have captured many statistical aspects of wind speeds very well, but is unrealistic in allowing changes between neighbouring states only.

It is interesting that no multivariate Markov-chain models for wind speeds were found. Given the advantages of Markov-chains, discussed in Chapter 3, building such a model (probably with a reduced number of zones) would appear to be a highly original contribution. Following Chapters will explore the possibility of combining multivariate Markov-chain and regression models into a single data generating mechanism.

This project does not favour hierarchical Markov-switching models, due to the large number of coefficients involved – and no overwhelming evidence was found that this model type should be adopted for hourly averaged wind. While Ailiot et al. [142] state that VAR models can only successfully describe the motions of meteorological systems that translate at a constant speed, this surely implies that a 3 hidden state model could then only describe 3 constant speeds. Hidden states clearly relate to other meteorological variables, perhaps circulation types, but this is not a field which has received much research attention yet.

With regard to modelling long memory, Bouette et al. [90] interpret a very significant peak in a wind speed periodogram at the annual frequency as a pole, and that the underlying process is GARMA. Caporin and Pres [152] fitted univariate ARFIMA-FIGARCH models, assuming that all seasonality is deterministic, but state they could have pursued alternatives such as GARMA or SARIMA models. This is a task they leave to other researches, this project clearly being among them. This research, it appears, can therefore be the first to apply a multivariate GARMA model to wind, and also the first to combine the Gagenbauer process with conditional heteroskedasticity.

Haslett and Raftery, fitting an ARFIMA model in [88] found that there is quite heavy long memory, with coefficient $d = 0.328$. They find the short memory aspect to be very light, which is somewhat surprising given the nature of purely short-memory models, but this could be a difference between daily rather and hourly series. Caporin & Pres working with Polish data found that long memory coefficients range from 0.17 – 0.2, a difference which could perhaps reflect weaker long memory in Poland, maybe due to a smaller influence of the NAOI. Alternatively the lower values could reflect the superior computational resources available to the authors and seem more consistent with the fact that short memory models can describe wind dynamics quite well.

Bouette et al. report that the estimation of a univariate GARMA model is very challenging, and the situation will of course be much worse with 20 dimensions. Fitting a GARMA model to one of the Irish sites used by Haslett and Raftery, they find the coefficient $\delta = 0.15$ from the periodogram's steepness near the pole, and $\delta = 0.18$ from the Whittle estimator.

Caporin and Pres [152] established that FIGARCH may be slightly better than GARCH for some zones, but not by much, suggesting that long-memory in variance is not an essential aspect of the model to be developed in this project. They fitted an ARFIMA model first, and then fitted a FIGARCH model to the residuals as a way of obtaining approximate values for both coefficient sets, before fitting all coefficients together. Given the additional complexity of the multivariate model, the separately fitted coefficients may have to suffice for the current project.

Some evidence was provided by Grothe and Schneiders [137] that conditional heteroskedasticity need not be modelled at all: initially Engle's ARCH test strongly indicated heteroskedasticity in residuals, but after deterministic detrending it did not. Tests will

obviously be carried out in this research on the residuals of the model for means, to determine whether a GARCH component to the model is necessary. GARCH volatility clustering can account for sudden changes in fluctuation behaviour, as described by Pinson and Madsen [143] for 10 minute averaged data.

There is much complexity associated with accurately converting single location wind speeds to zonal power outputs, leading Hill et al. [132] to avoid attempting it.

For a very simplified approach, tables are available from previous work at Bath listing the capacity in different location types (coastal, offshore, highland and lowland) in each zone for a choice of scenarios. These can be up-scaled and adjusted based on capacity that definitely has been, is being and definitely will be built. Tables are also available providing information on how the Bath Model's set of Met Office stations should be re-scaled for each location type. These can be adapted to reflect any choice of Met Office stations. Ideally, final hub-height wind speeds should be transformed to having site-accurate Weibull shape factors, but sufficient data are not available.

Following the Bath model, the number of turbines in each location type in each zone will be calculated and their availability assumed independent Bernoulli distributed. Ideally the sophistication should be added of assuming that if an offshore turbine becomes unavailable in winter, it should remain so until the spring. This project may benefit from turbine power curves provided by Garrad Hassan Ltd which account for wake and similar losses. Manufacturers also provide power curves that account for the smoothing effect of multiple turbines, as used by Gibescu et al. in [146]. Nørgaard and Holttinen [147] provide a fairly simple methodology for smoothing wind speeds to represent a region of a given size. Holttinen [55] also presents useful principles for the relationship between standard deviation and mean for areas of different sizes.

All of the studies described here focus on the recreation of some aspect of the wind resource's dynamics in detail, such as getting the multivariate relationship correct, the long memory, diurnal seasonality or the smoothing effects of multiple turbines. None consider all such aspects with much sophistication since complexity renders this unfeasible, but this project will seek compromises that give at least reasonably good consideration to all relevant aspects.

Chapter 5

Data Acquisition and Processing

This chapter describes and presents results of the initial research stage of the project. This includes: how historical wind speed datasets were sourced and converted into a useful form; the algorithm developed for correcting chronological errors in the data; an investigation into the possibility of identifying false zero wind speed recordings; the process of choosing the optimum meteorological stations; an analysis of the quality of mast locations for the chosen stations; the process of transforming the series to be approximately stationary with a multivariate normal distribution – including identification and removal of diurnal seasonality in both mean and standard deviation; and comparison of the performance of two algorithms for filling-in missing data.

5.1. Initial Data Acquisition

The wind models developed for this research project are applied to historical hourly averaged wind speed data recorded by the UK's Meteorological Office. Since the Bath Wind Model (BWM) is the starting point, one observation station is required within each BWM zone. This data is made available to academic researchers through the website of the British Atmospheric Data Centre (BADC) [153], contained in a database named Midas. Data was downloaded for 3 sites: Lerwick, Leeming and Capel Curig, and viewed/ manipulated in *Microsoft Excel*. The datasets contained many variables of no relevance to this research, which were deleted to leave only the average wind speed and the time stamp. Although not used in the current project, maximum gust speeds are available and could be useful for future analysis, providing a very rough idea of the extent of turbulence during the hour in question. The wind speeds are rounded up to the nearest knot (nautical miles per hour), equivalent to 0.514m/s. Table 5.1 gives an example of a short segment from one of the series, with data quality slightly below average.

Time Stamp	Wind Speed (Knots)
2005-11-07 14:58	26
2005-11-07 15:59	27
2005-11-07 16:58	29
2005-11-08 08:51	28
2005-11-08 08:52	31
2005-11-07 20:11	
2005-11-07 20:59	37
2005-11-07 22:36	34
2005-11-08 00:38	33
2005-11-08 00:43	24
2005-11-08 00:58	18
2005-11-08 01:58	15
2005-11-08 02:58	15

Table 5.1. Example Segment of the raw Met Office time series

Initially, every available hour for the period 1st Jan 2005 – 31st Dec 2007 was downloaded for the three sites, followed by a further download for the period 1st Jan 1988 – 31st Dec 2004 at Lerwick. This station is at a windy location on the Shetland Islands to the north of Scotland (Lat: 60.133^{o}, Long: -1.183^{o}, Alt: 82m), whilst Leeming is located just to the east of the Pennine hills in North Yorkshire (Lat: 54.3^{o}, Long: -1.533^{o}, Alt: 32m) and Capel Curig lies in central Snowdonia, north Wales (Lat: 53.1^{o}, Long: -3.933^{o}, Alt: 216m). The first two sites were identified by Miranda and Dunn in [25] as having high quality data, while the 3rd was picked at random by the Author, based on personal familiarity, in order to gain some idea of how typical is the data quality at the first two sites.

All analysis and model fitting work involved in this project was conducted using the language/ software package *Matlab*. Code was written to read the data and store it in the desirable form – each hour a vector of numbers: wind speeds, year, month, day of the month, and hour of the day. The total number of hours passed since an arbitrary point in the past (1st of Jan 1950) was added to act as a time-stamp index. This means that the time series were represented as matrices, with all elements purely numerical, the most natural type of object to me manipulated in *Matlab*.

Despite knots being non-standard units, it was decided to keep them in this form, since converting to m/s would surrender the coding advantages of working with integers. Rounding to the integer m/s, a scale with roughly half the resolution, would involve significant loss of information. The data was found to contain many different types of errors including missing wind speeds, multiple entries for the same hour, missing hours and incorrect chronology. Several examples of such errors can be seen in Table 5.1.

Figure 5.1 shows a typical time series segment, covering a period of 10 days. First visual impressions confirm that most inter-hourly changes are small, but the wind can change a great deal in a few hours, and that patterns display a degree of self-similarity on different time scales. This is confirmed by plots of the time series when compacted into daily averages. Deterministic patterns, whist known to be present, do not seem easily identifiable from the stochastic behaviour.

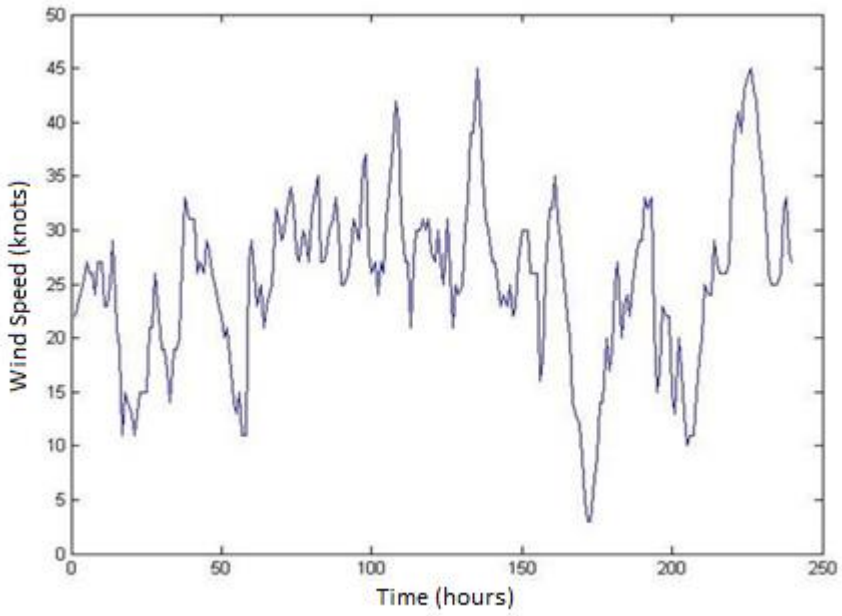

Fig. 5.1. Wind speed time series plot, Lerwick, 1st 10 days of Jan 2005.

5.2. Cleaning-Up the Datasets

5.2.1. The Clean-Up Algorithm

An algorithm was developed to 'clean-up' the data to ensure correct chronology and that the most likely of multiple entries is chosen. The components of this algorithm were:

1) Allocate the deliberately impossible wind speed of -100 knots For all entries with a time stamp but no wind speed recording – this is the entry for all invalid wind speeds, and keeps the matrices purely numerical. Some nonsensical wind speeds were found in the data, i.e. negative of too large. These were converted to -100 knots, with 200 knots arbitrarily chosen as the threshold for high wind.

2) Create a 6 x ls matrix M, where ls is the length of the 'raw' series. The 1st row must consist of only -100 repeated; rows 2 – 5 consisting of the time stamps broken down into year, month, day of month and hour of month; and the 6th row consisting of the hour index described above;
3) Moving left to right (i.e. forward in time), for all columns i where $M(6, i+1) - M(6, i) = 1$ (i.e. columns where the time stamp is exactly one hour ahead of the previous column) replace $M(1, i)$ and $M(1, i+1)$ with the i^{th} and $(i+1)^{\text{th}}$ wind speeds. The resulting 1st row is considered to represent a tentative set of valid wind speeds, a series of blocks in which there is correct chronological progression. The process inevitably involves the discarding of some accurate recordings, the underlying problem being that we cannot know exactly which ones are accurate.
4) Create a list of columns where valid blocks end, and the corresponding values for $\Delta t_j = M(6, j+1) - M(6, j)$. Where $\Delta t > 0$, chronological order is correct, but data is missing. Where $\Delta t < 0$, chronological order is incorrect and there is clearly a block which does not belong, which must be deleted. Chronological progression in the correct direction may be restored by either removing the block beginning at the junction where $\Delta t < 0$ or the block ending at it.
5) Move left to right through the matrix, deciding which of the above is the case for each problematic junction and remove blocks accordingly (i.e. delete the entire columns within the offending block). Specific code is required to ensure that the last pair of blocks are chronologically correct. This process does not guarantee that the entire time series has correct chronological order, some combinations of errors can 'slip through'.

6) Manually check for any remaining error, and correct as necessary. All datasets cleaned in the current research project were examined by human in detail, with no further need for manipulatio identified.
7) Create a new matrix M_{Full}, with 6 rows and a number of columns equal to the number of hours in the time range from the 1st to the last valid wind speed in M. Fill rows 2 – 6 with the correct time-stamps and index values. Take all valid wind speeds from M and place them in the correct positions in the 1st row of M_{Full}. Fill the remaining columns of the 1st row of M_{Full} with -100.

It was found later that data amounts discarded by the algorithm are rarely more than 1 – 2%.

5.2.2. Excess Zeros and Fitting Weibull Distributions

With the data input and clean-up algorithms developed, it was possible to explore the time series at many Met Office stations across GB, with missing data ignored. The simplest means of examination was to plot wind speed histograms. Figure 5.2 shows the histogram for Lerwick, displaying an almost perfect Weibull distribution.

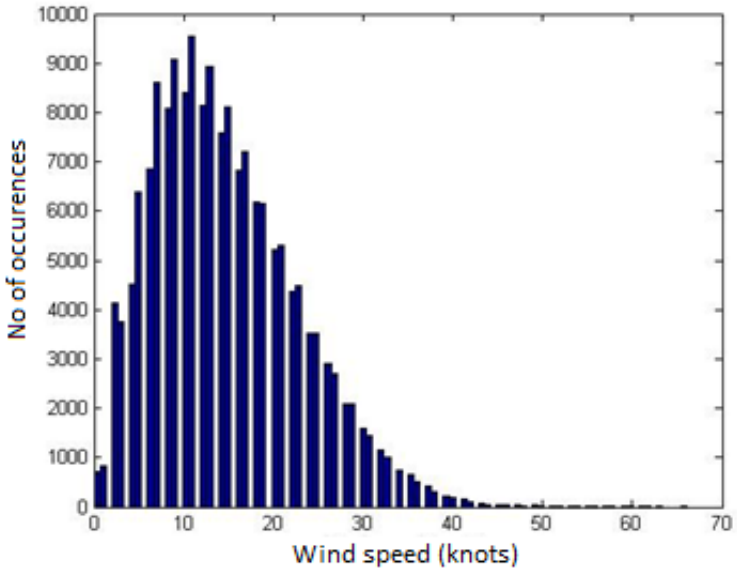

Fig. 5.2. Wind speed histogram for Lerwick, 1988 - 2007.

Some stations, including Machrihanish in south western Scotland, displayed anomalies from the Weibull distribution such as a very excessive number of readings of a specific, low speed, as shown in Figure 5.3 below.

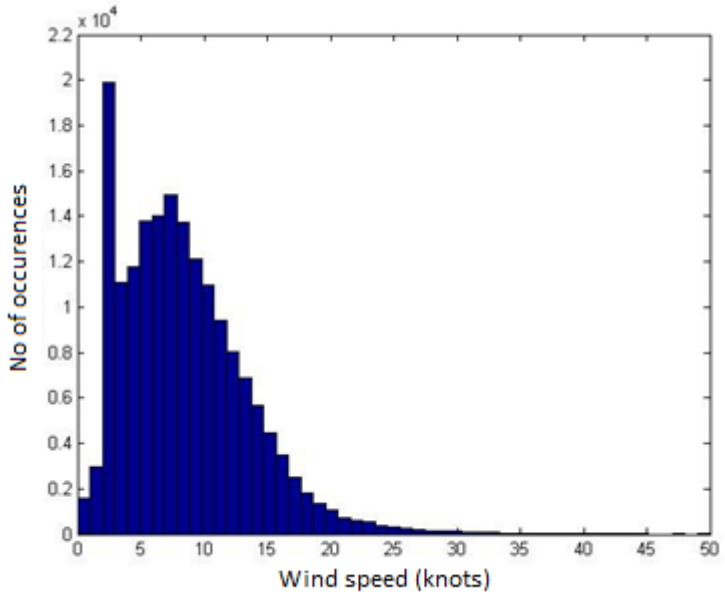

Fig. 5.3. Wind speed histogram for Machrihanish, 1988 - 2007.

A more common deviation from the distribution, found at the majority of stations, is an excess of zero knot recordings – as anticipated in Chapter 2. A user guide for the Midas dataset [154] found on the BADC website confirms that measurements have historically been made using a cup anemometer, which is a source of error for low wind speeds, due to friction. This is clearly the reason for the excess zeros (and possibly the excess 2 knot readings), along with periods during which the anemometer has frozen.

To examine the situation more precisely, the best fit Weibull parameters for the various station distributions were approximated. This was achieved by estimating the reverse CDF's for the series, $\widehat{Q_W}(u; k_W, C_W)$, where u is wind speed, and making use of

$$Q_W(u; k_W, C_W) = \exp\left(-(u/C_W)^{kw}\right), \tag{5.1}$$

so that $ln(ln(\hat{Q}_W))$ vs. $ln(u)$ is a straight line with gradient $\widehat{k_W}$ and intercept – $\widehat{k_W}/\widehat{C_W}$. It was found that all stations produced graphs that were very nearly straight lines, as shown for Leeming in Figure 5.4. For Leeming, the parameters were $\hat{k}_W = 1.5$ and $\widehat{C_W} = 8.47$.

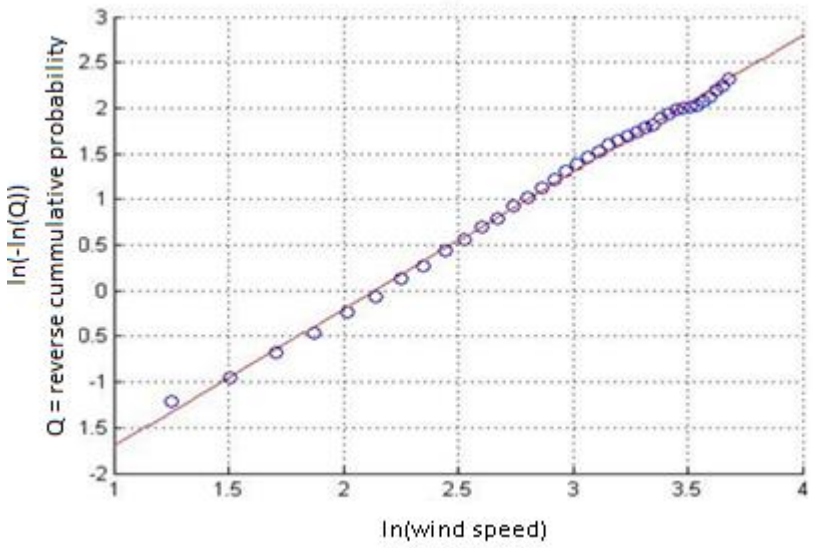

Fig. 5.4. ln(ln(CDF)) vs. ln(wind speed) for Leeming, 1988 - 2007.
Straight line implies Weibull distribution.

Having established the best fit parameters, corresponding density functions were plotted along with normalised histograms. The result for Leeming, shown in Figure 5.5, was found to be typical i.e. excess zeros at the expense of low wind speeds, then a good fit. It is possible that a mixed Gaussian distribution would be a better fit for a minority of stations. The deviations from Weibull for low wind speeds appear much more extreme when looking at pdf's rather than cdf's.

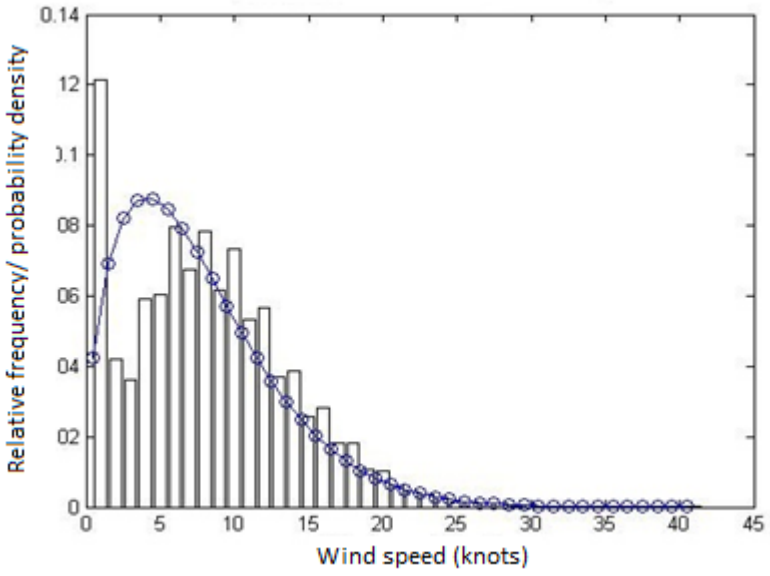

Fig. 5.5. Normalised histogram for Leeming, 1988 – 2007,
& Weibull distribution with best fit parameters.

As described in Chapter 2, in such cases one may assume that the probability density is a Dirac delta function (at zero) and a Weibull distribution for speeds > 0.

5.2.3. Examination of Zero Wind Speed Chains

Although there is nothing that can be done to accurately counter the effects of friction, it may be the case that an additional step to the clean-up algorithm could be developed to identify and remove erroneous zeros caused by the anemometer either freezing or being unable to rotate for some other reason. It seems reasonable that such periods might be identifiable as a series of consecutive zeros of lengths distinctly greater than those arising from true recordings of calm periods. Fig. 5.6 shows a histogram of the length of consecutive zeros found in the Capel Curig data, before it was subjected to the clean-up algorithm. The histogram indicates than in fact one cannot observe two distinct categories of consecutive zeros in the data. One may say that the few chains of greater than ~ 30 consecutive zeros are probably outliers, but in fact the clean-up algorithm eliminated most of these.

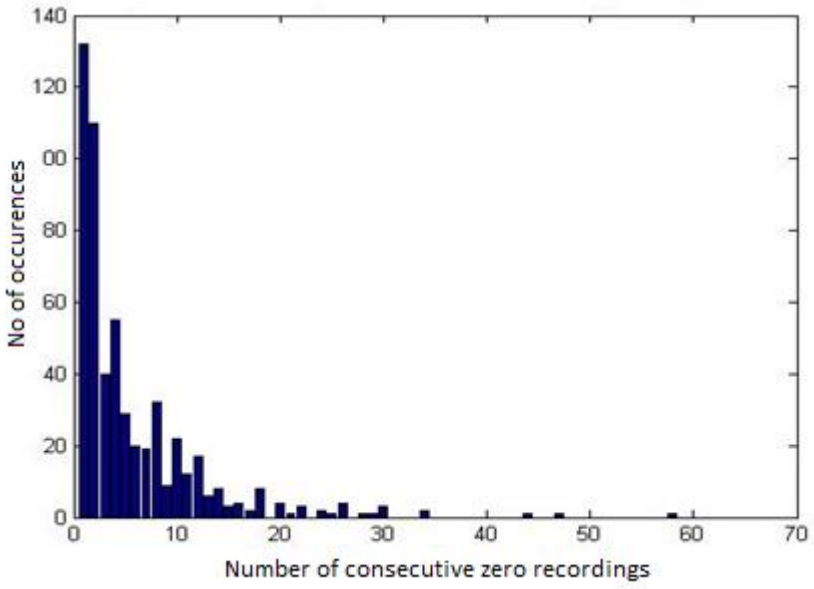

Fig. 5.6. Histogram of the length of consecutive zero wind speed recordings, for raw data, Capel Curig. sample.

To be certain, additional analysis was carried out, seen in Fig. 5.6, in which the total accumulated number of specific wind speed recordings are viewed as a function of consecutive chain length – for zero wind speed and two other (arbitrary) wind speeds. The figure shows

that chains with lengths across the range from 10 – 30 all contribute to the excess zero count, not present for the other numbers, but there is no basis for choosing a particular value within this range as a cut-off point between 'true' and 'frozen/ stuck' zero chains. It was seen that this was the case for several Meteorological stations across GB.

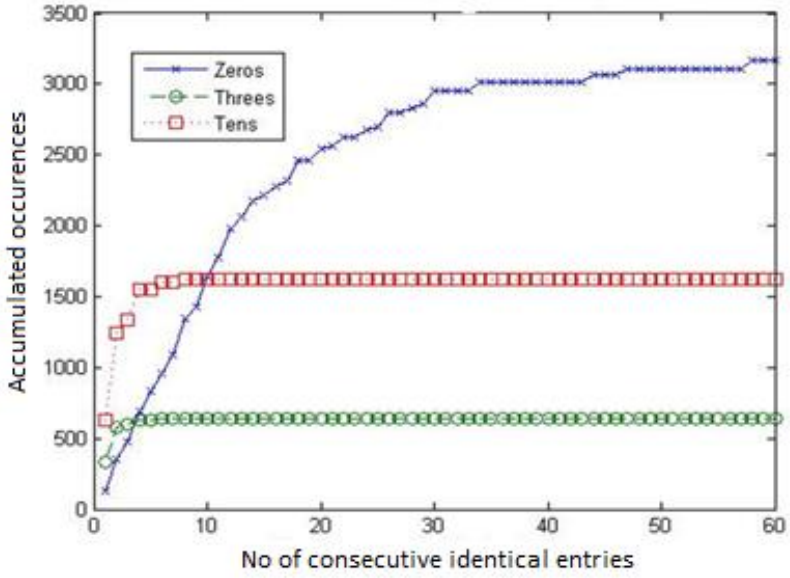

Fig. 5.7. Total accumulated occurrences of specific wind speed recordings, 'cleaned' Capel Curig sample.

Another potential source of excess zeros is failure of the data logging equipment. Such failure might be characterised by a sudden drop in the recorded wind speed from its true value to zero, so attempts were made to identify such patterns in the data. Histograms were once more produced for the change in wind speed occurring just before the first zero in a chain, seen in Fig. 5.8. For comparison this was also done for other arbitrary wind speeds, and 10 knots is shown Fig. 5.9. The histograms are in fact very similar, although zero recordings are in fact preceded by relatively fewer changes of greater than 5 knots, and there definitely is no justifiable cut-off point for distinguishing between true and false recordings. The conclusion drawn therefore was that nothing should be done to remove any of the excess zero recordings, and that the clean-up algorithm is therefore complete. This was checked with several stations to ensure generality.

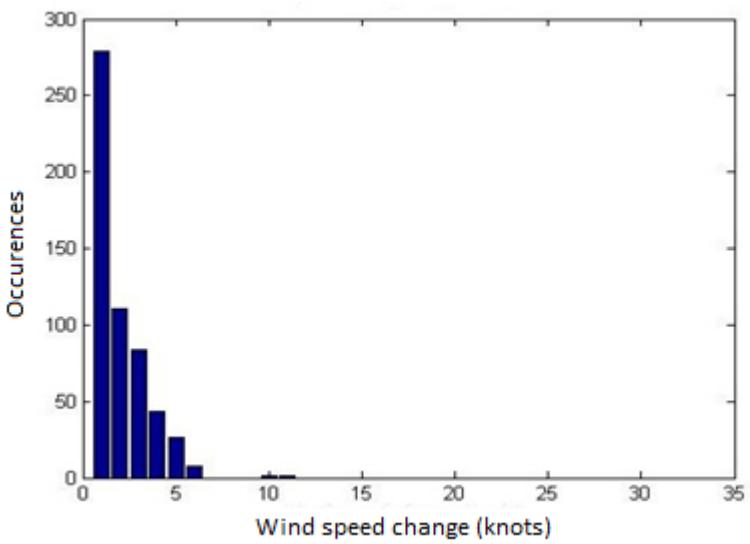

Fig. 5.8. Histogram of wind speed changes occurring prior to recordings of zero wind speed, 'cleaned' Capel Curig sample.

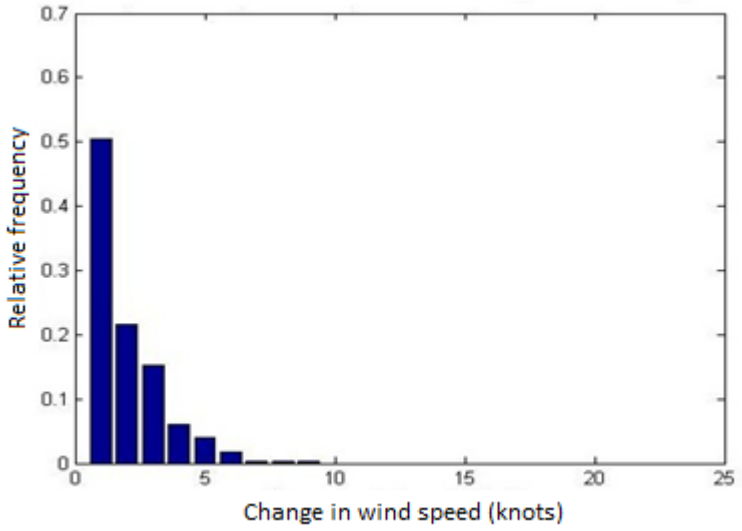

Fig. 5.9. Histogram of (absolute) wind speed changes occurring prior to recordings of 10 knots, 'cleaned' Capel Curig sample.

5.3. Choice of Meteorological Stations

5.3.1. Choosing the Stations

With the process of cleaning-up the raw time series optimised, the next task was to decide upon the best set of 20 weather stations for representing the wind speed field across GB. It was essential that the choice of stations ensured:

- Good, even coverage of GB, including all geographical extremities;
- There is a station close to all major development areas for wind capacity;
- The stations are in reasonably simple, open terrain;
- None of the station masts are adversely affected by nearby obstacles;
- All stations have a very high % of data present, as a whole sample and for each sub-sample of reasonable length (e.g. 5 years); and
- At all stations, gaps of more than 2-3 days are rare, and that there are none longer than a few weeks.

Since a major goal of this project is the accurate simulation of very low frequency variability, it is advantageous for the historical datasets to be as long as possible. The investigations described so far involved 20 year samples, from 1st of Jan 1988 to 31 December 2007, and the first task in choosing stations was to see whether this could be extended to e.g. 25 or 30 years. Investigations involving stations chosen for the previous Bath Wind Model work revealed that it was unlikely that the period could be extended, while satisfying the requirements listed above, and that the years 1988 – 2007 are as good a choice as any. The task was therefore to find the set of sites that satisfy the requirements above, to the greatest extent possible, for this 20 year period. Even at 20 years, it was decided that the sample size is so large that no part needed to be reserved for out-of-sample validation.

Selection proceeded by using the BADC website to find as many meteorological stations as possible that appear to be reasonably well located, in terms of the requirements listed above, and subject them to analysis. Initial checks on a potential location were the percentages of data present in the whole sample, and in each 5 year sub-sample, and also the distribution of lengths of missing data periods. It became evident that while the clean-up algorithm was always successful in securing correct temporal order, on some occasions where the time series contain two consecutive entries for the same hour, the block of data starting at this point and extending up to the next discontinuity is erroneously marked as invalid – and such blocks may be long. As a result, for each station analysed, all large gaps had to be manually checked in case they were in fact present due to this fault. Having completed the

analysis on gaps, many combinations of stations were considered, with slightly lower data quality potentially accepted if e.g. the overall evenness of geographical spread was improved. The final list of station names and their overall % of data present for the 20 year sample is given in Table 5.2 below. The choice of stations made for the previous Bath modelling work was maintained for 11 of the zones.

Zone	Station Name	Percentage Present
1	Lerwick	99.60
2	Stornoway Airport	98.89
3	Wick Airport	94.93
4	Tiree	98.41
5	Peterhead Harbour	94.92
6	Dunstaffnage	96.28
7	Machrihanish	98.71
8	Salsburgh	94.60
9	West Freugh	99.40
10	Carlisle	96.12
11	Valley	99.87
12	Leeming	99.76
13	Waddington	99.92
14	Aberporth	99.37
15	Shawbury	99.33
16	Marham	99.41
17	Camborne	99.80
18	Solent	95.50
19	Northolt	95.41
20	Manston	98.70

Table 5.2. Meteorological station names for each GB Zone, and percentages of data present.

The selection process was successful, with the lowest value 94.6%, and 9 stations having more than 99% of data present. Further, it was found that the percentage of hours for which data is present for all 20 locations is 66.35%. This was close to the value of 66.11% obtained by assuming the errors at different locations to be uncorrelated.

5.3.2. Quality of the Chosen Locations

The locations of the chosen stations, along with zone boundaries, are shown in Figure 5.10 below. In order to assess the proximity of the stations to areas of future high wind capacity, Figure 5.11 shows the locations of all GB wind-farms currently operational, under construction or consented. This is taken (with minor adjustments) from the wind energy database section of the website of RenewableUK, the trade and professional body for the UK wind industry [2]. It seems reasonable to assume that the spatial distribution of onshore capacity up to a few decades into the future will not differ significantly from this pattern. There are a total of 669 projects on the map, some of which are clearly too close together to be resolved on the scale shown.

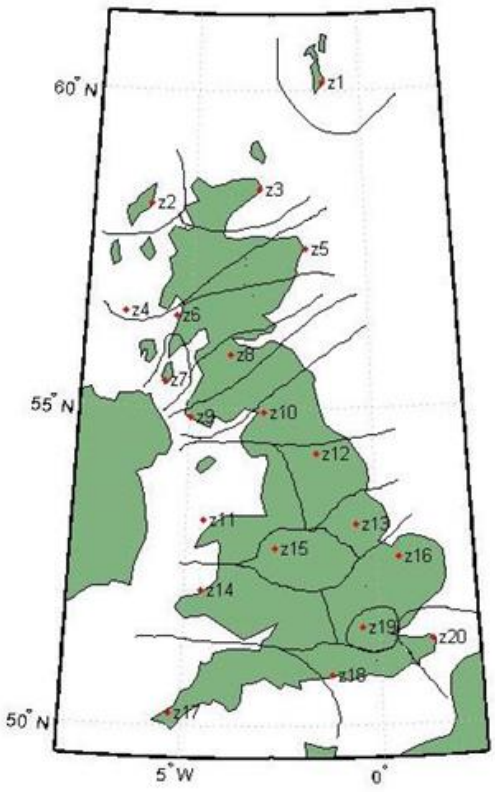

Fig 5.10. The locations of chosen stations, with zone boundaries.

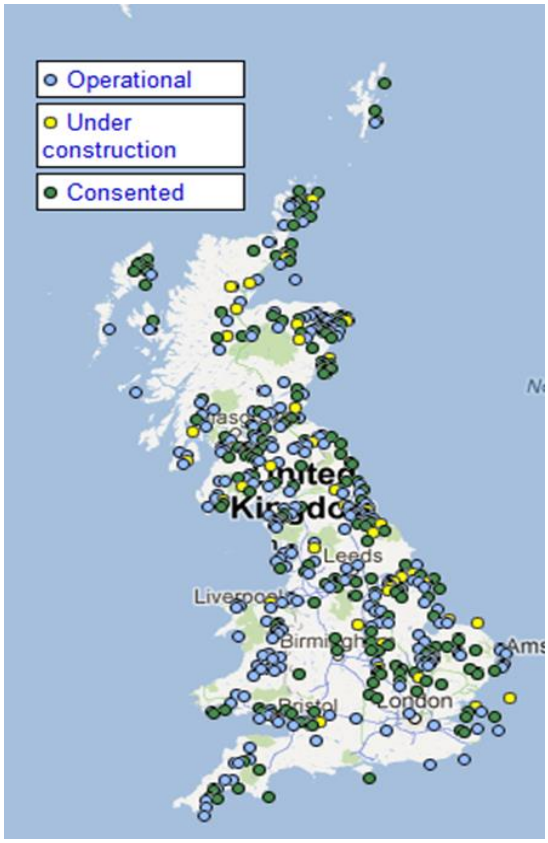

Fig 5.11. The locations of wind-farms in GB that were operational, under construction or consented, 04/2010.

Figure 5.10 shows that the stations give a reasonably even coverage of GB. The positions of stations in Zones 1 and 2 could not be much different and fortunately the station in Zone 2 is very close to a number of wind-farms. The station for Zone 3 is also located very close to a significant cluster of generators, including some on the Orkney Islands. The same is true for Zone 5, with a very large number of projects running from Peterhead east towards Inverness and south towards Aberdeen. The stations for Zones 4 and 6 are unfortunately close together, and neither is close to significant wind turbine capacity. The situation would be vastly improved if the station for Zone 4 were re-positioned about 100 - 150km to the south east, and the station for Zone 6 were re-positioned near the southern edge of the zone, close to the cluster of projects in the Scottish central lowlands. This was not possible due to a severe lack of meteorological stations with wind records in this part of Scotland, and the sites chosen were the only two available with reasonable data quality.

The choice for Zone 7 was limited by spatial configuration, but there is some capacity within a moderate distance. Zones 8, 9, 10, 12, 13 and 16 all have well located stations both in terms of proximity to capacity and being evenly spread out. The station for Zone 11 might be closer to more wind-farms if it were located in the Morecambe Bay area, but it would then be

quite close to the stations for Zones 10 and 12, so the chosen location on the Isle of Anglesey gives a better coverage of GB, along with excellent data quality. The station for zone 15 would be closer to more capacity if it were positioned towards the east, rather than west of the zone, but would then be too close to the stations for Zones 13 and 16. The station for Zone 14 is fairly well located for the wind capacity in the hills of mid Wales, but would be better if it were positioned between this cluster and the capacity stretching along the south Wales coast and the Bristol Channel. Again, it was a lack of suitable stations which prevented this. It was fortunate that for Zone 17 a suitable station was available in western Cornwall, close to capacity there and extending the total geographical reach of the stations. The stations for Zones 18 and 19 are not close to any significant capacity, but this is inevitable given the very small number of projects in these (densely populated) zones, and the station choices achieves the best possible even spread for south east England. Choices were limited for Zone 20, but the chosen station is close to significant capacity in the Thames Estuary.

5.3.3. Description of the Precise Mast Locations

A brief description is now given of each meteorological station. The narrative descriptions are based on observation of the locations on Google Earth, given the high precision values for longitude and latitude. The level of precision is unfortunately not enough to know exactly where the masts lie in relation to buildings, and this uncertainty is reflected in the narrative below where appropriate.

Zone 1, Lerwick. Latitude: 60.1395, longitude: -1.1829, altitude: 82m, station id: 9. This station is located in a very exposed position on a small plateau, close to the sea, on the main island of Shetland. There are no buildings or other obstructions in the vicinity. This is the most northerly station.

Zone 2, Stornoway Airport. Latitude: 58.2138, longitude: -6.3177, altitude: 15m, station id: 54. The station is located in a very flat, open landscape on the eastern side of the Isle of Lewis. The location is only a few metres from the sea, but might be slightly sheltered from westerly/ south-westerly winds by hills on the western side of the island. There are a few small airport buildings in the vicinity, which might cause very minor sheltering effects.

Zone 3, Wick Airport. Latitude: 58.4541, longitude: -3.0884, altitude: 38m, station id: 32. The station is again surrounded by a very flat landscape, about 2km from the sea. There is a possibility of some sheltering effects from airport buildings.

Zone 4, Tiree. Latitude: 56.5, longitude: -6.8796, altitude: 9m, station id: 18974. The station is located on the small inner Hebridean island of Tiree. The island is very flat, so that the station is completely exposed to winds from all directions, including Atlantic storms. There are no buildings in the vicinity. This is the most westerly located station.

Zone 5, Peterhead Harbour. Latitude: 57.5027, longitude: -1.7726, altitude: 15m, station id: 170. Once again this station is very close to the sea and is surrounded by a very flat landscape. There are several harbour buildings and large machinery in the vicinity which may affect the wind speeds somewhat.

Zone 6, Dunstaffnage. Latitude: 56.4505, longitude: -5.4386, altitude: 3m, station id: 918. While the vicinity of this station is flat and free of obstacles, there are nearby hills in most directions, some of them steep. The station is only a few metres from the sea, although the Isle of Mull lies between it and the open ocean.

Zone 7, Machrihanish. Latitude: 55.4408, longitude: -5.6957, altitude: 10m, station id: 908. This is another station based at a small airport. On the western edge of the Kintyre Peninsula, about 1km from the sea, the location is quite exposed. The landscape is very flat for a few km in each direction, but this area is surrounded by moderately complex terrain. There are few if any obstacles in the immediate vicinity.

Zone 8, Salsburgh. Latitude: 55.8615, longitude: -3.8754, altitude: 277m, station id: 982. This site is inland, at the centre of the Scottish central lowlands, but has significantly greater altitude than any other site. The terrain is open, with gentle hills. The location is surrounded by agricultural land, although there are some buildings is the vicinity which may possibly have some influence.

Zone 9, West Freugh. Latitude: 54.859, longitude: -4.9341, altitude: 11m, station id: 1039. This is yet another small airport location. The terrain is flat and exposed, although there are some fairly gentle hills about 5km away, and the site is about 3km from the sea. The precise location appears to be too far from airport buildings for them to have any effect.

Zone 10, Carlisle. Latitude: 54.9342, longitude: -2.9622, altitude: 28m, station id: 1070. This station is located on the northern outskirts of Carlisle, near what appears to be a business park. The site is inland, and the terrain flat for many km in every direction. The scattered buildings of the business park may cause some mild sheltering effects.

Zone 11, Valley. Latitude: 53.2524, longitude: -4.5352, altitude: 10m, station id: 1145. This station is located at RAF Valley on the Isle of Anglesey. The surrounding landscape is flat in every direction for a significant distance, and the site is 1-2km from the Irish Sea. There might be a few airport buildings in the vicinity of the mast.

Zone 12, Leeming. Latitude: 54.2968, longitude: -1.5315, altitude: 33m, station id: 17314. Another station located at an airbase, surrounded by flat open countryside and a village. Some buildings in the village might have minor impact when the wind is blowing from the south.

Zone 13, Waddington. Latitude: 53.1751, longitude: -0.5217, altitude: 68m, station id: 384. This site fits exactly the description of Leeming above, except that the village of Waddington is west of the mast.

Zone 14, Aberporth. Latitude: 52.1391, longitude: -4.57, altitude: 133m, station id: 1198. Another airbase, but on this occasion near the edge of a sea cliff, overlooking the southern end of the Irish Sea – so somewhat exposed to Atlantic storms. The surrounding countryside is not flat, but the hills are gentle. There are only a few small buildings in the vicinity of the mast.

Zone 15, Shawbury. Latitude: 52.7943, longitude: -2.6633, altitude: 72m, station id: 643. Airbase, surrounded by quite flat countryside, with a few gentle hills. Again, there is a possibility that airfield buildings and the village of Shawbury might cause minor sheltering effects.

Zone 16, Marham. Latitude: 52.651, longitude: 0.5677, altitude: 21m, station id: 409. Another airfield, surrounded by completely flat countryside. The mast appears to be a reasonable distance away from the airfield buildings and the village of Marham.

Zone 17, Camborne. Latitude: 50.2178, longitude: -5.3266, altitude: 87m, station id: 1395. This station is located in the middle of agricultural land, in a landscape of gentle rolling hills, about 10km from the Atlantic coast of Cornwall and close to the town of Camborne. There are a few buildings scattered about, but it is unlikely that any are directly sheltering the mast. This is the most southern of the locations.

Zone 18, Solent. Latitude: 50.8075, longitude: -1.2092, altitude: 9m, station id: 858. This station appears to be located in or adjacent to a hovercraft depot next to an airfield, only a few metres from the English Channel. The landscape is flat. The vicinity is built-up, with quite a few potentially sheltering buildings/ objects – but the extent of their impact is impossible to know.

Zone 19, Northolt. Latitude: 51.5453, longitude: -0.4153, altitude: 33m, station id: 709. Another airfield, in a flat landscape, within Greater London. There are some fairly tall buildings which could potentially have some sheltering effect.

Zone 20, Manston. Latitude: 51.346, longitude: 1.3372, altitude: 49m, station id: 775. An airfield in a flat landscape, at the eastern most tip of southern England, it is a few km away from both the mouth of the Thames Estuary and the English Channel. It is the most eastern of the stations. Airfield buildings appear to be a reasonable distance away from the mast.

To summarise, the locations are generally good – some of them ideal, and the selection is probably as good as can be reasonably expected, given the numerous simultaneous demands to be met by the choice of stations.

5.4. Filling-in Missing Data

While it was possible to find stations with very high percentages of data present, calculations involving *Matlab* may occasionally require the wind speed and time stamp data array to be 100% filled with reasonable values. Such complete matrices are also useful since many time series related functions are not general enough, and had to be constructed from much more general ones. This lead to data gaps being filled, despite the fact that statistical inferences would be more accurate if this were not the case. Filling the gaps in the best possible way represented the next major challenge for the project. Before appropriate techniques could be developed and tested, some processing of the series was necessary, and this is described in the following sections.

5.4.1. Power Transformation and Standardisation of the Wind Speeds

Part of the data gap-filling process will rely on the assumption that the series is Gaussian. It was established in Section 5.2.2 above that the assumption of a Weibull distribution for the series is reasonable, despite the problems at low wind speeds. Power transformation will therefore be used, followed by subtraction of means and division by the standard deviations, to create an approximate MVN distribution, with marginal $N(0,1)$ distributions. These transformations are also a pre-requisite for the modelling techniques used later.

A viable alternative would be to take logs of the wind speeds, following the example of Caporin and Pres [152], described in Chapter 4. This approach has the advantage of probably linearising the process in a way that power transformation cannot, and also the transformed process would be supported for all $\mathbb{R}$, rather than only $\mathbb{R}^+$. However since the log transformation would change the series to a greater extent than a power transformation, modelling errors would probably become more significant upon reversal of the transformation, leading to a less accurate final output. Also, the approach taken here is to allow the simulated process to take negative values, and convert them to zero wind speed as a final step, under the very reasonable assumption that the proportion of negative values will be very small and the resulting excess zeros will be fewer than found in the historical data.

As discussed in Chapter 4, Section 4.2.1, a Weibull distribution with a shape parameter of 3.6 is very close to the Gaussian. If a series distributed with shape parameter k_W is raised to the power δ_W then the new shape parameter will be k/δ_W. Therefore, the series can be rendered approximately Gaussian by raising them to the power $\delta_W = \widehat{k_W}/3.6$, where the $\widehat{k_W}$'s

are estimates obtained using the method described in Section 5.2.2 above. These estimates are unlikely to be exactly correct, so several values close to them were tried, to see which one gives the most Gaussian-like result.

In the multivariate case, it is desirable to find a compromised common value for δ_W which transforms all zones to be Gaussian to a reasonable approximation. This is quite challenging with 20 zones, and it would be easier if a sub-set were identified as more important, in some sense. A natural way to do this is to select the zones in which the greatest amount of wind capacity is present or, better, the zones where the greatest amount of capacity will be present in about a decade. Previous Bath modelling work [27] developed two scenarios for the zonal distribution of wind capacities in 2020 - one which favours strong onshore development in Scotland, the other favouring England and Wales, with a higher percentage offshore. Zones were arranged according to their capacity in these scenarios, with the highest capacity zone first. If certain zones appear towards the beginning of both lists, despite the considerable differences between the scenarios, then this suggests that it is highly likely that large wind capacities will be present in those zones.

For the Scotland centred scenario, the order is 8, 6, 3, 11, 2, 9, 16, 5, 12, 20, 14, 4, 10, 1, 7, 13, 17, 15, 19, 18; with 46% of capacity concentrated in the 1st five zones. For the other scenario the order is 11, 16, 12, 20, 13, 10, 14, 5, 8, 9, 6, 4, 3, 17, 7, 18, 15, 19, 2, 1; with 64% of capacity concentrated in the 1st 5 zones. Consideration of these rankings, combined with examination of Figure 5.11, lead to the conclusion that the 5 most important zones are 5, 8, 9, 12 and 16. The range of values for $\hat{k}_W$ for these zones suggested that δ_W is in the range 0.4 – 0.5. Investigation showed that 0.4 is close to optimal for Zone 8, pretty good for Zone 5, and is generally the best all-round choice. After removal of means and division by standard deviations, it was found that the covariance matrix satisfied the criterion of being positive-definite so that the series could be reasonably modelled as a MVN.

5.4.2. Removing the Diurnal Seasonality

It was stated in Chapter 4 that since we are fitting only one model, capable of simulating the behaviour of wind throughout the year, and that it is unlikely that the diurnal seasonality can be modelled as stochastic in nature, since the simulated process would not reproduce the manner in which the diurnal seasonality changes throughout the year. The diurnal seasonality must therefore be assumed deterministic in nature. The next stage in

processing the historical series is the identification and removal of this seasonality. An algorithm was developed to identify the seasonality which worked as follows:

- For each hour, t, calculate the difference between the wind speed and the daily 'background' wind conditions – an equally weighted moving average of all hours in the period $t - 11$ to $t + 11$;
- For each month and hour of the day, calculate monthly diurnal profiles: the mean values of this difference from background; and
- Create diurnal profiles for each day of the year by assuming the monthly profiles represent the daily profile at the mid-point of each month, and interpolating linearly between them.
- Do this separately for leap and normal years.

All zones revealed profiles of a similar nature, and much more significant in summer than winter in each case. The phenomenon is considerably more prominent in some zones, and some profiles are more symmetric than others about the time of peak windiness – generally mid-afternoon. It is interesting to note that the diurnal trend at Salsburgh (Zone 8), inland, is very similar to that at Machrihanish (Zone 7) on the coast – i.e. although the effect is due to differential heating of the land and sea, it is not only limited to the coast. The strongest effect by far is for Peterhead Harbour (Zone 5), with Lerwick (Zone 1) second – both facing the North Sea.

The methodology was validated by splitting the 20-year samples into 2 equal subsets, applying the algorithm to each, and examining how similar the derived difference profiles are. Figure 5.12 shows the results for 2 contrasting months for Lerwick. The profiles for July are very similar, January not as much so, but this isn't very important since neither 'versions' of it involve a large effect. As further validation, the autocorrelation function was examined for the range of lags 20 to 1000 hours, before and after diurnal detrending – shown for Lerwick in Figures 5.13 and 5.14.

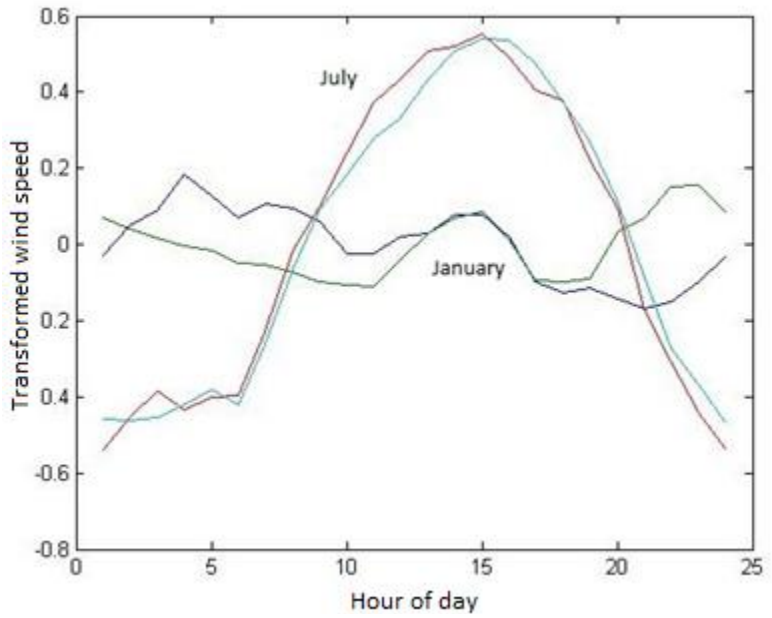

Fig. 5.12. The diurnal difference profiles for Lerwick, two 10 year samples.

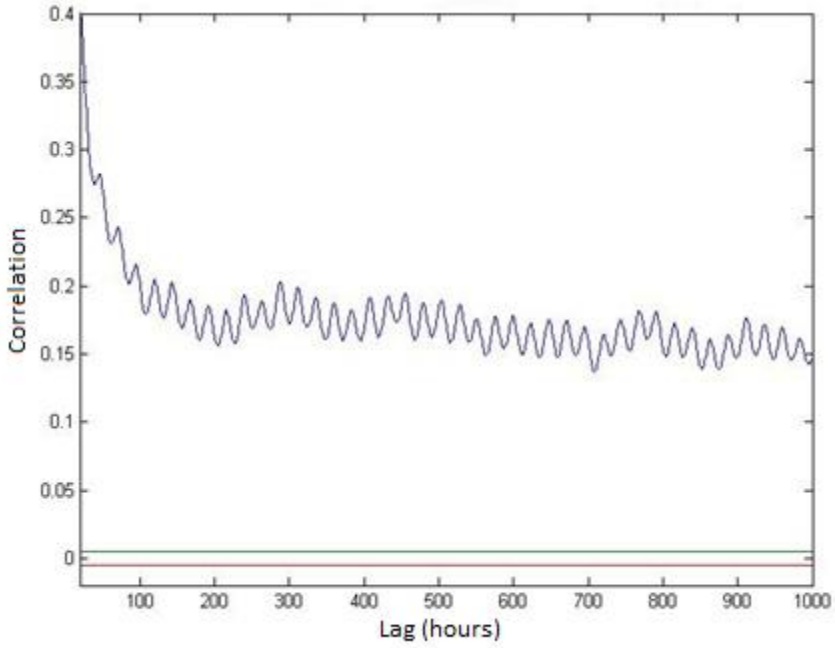

Fig. 5.13. The ACF for Lerwick, power transformed, lags 20-1000 hours.

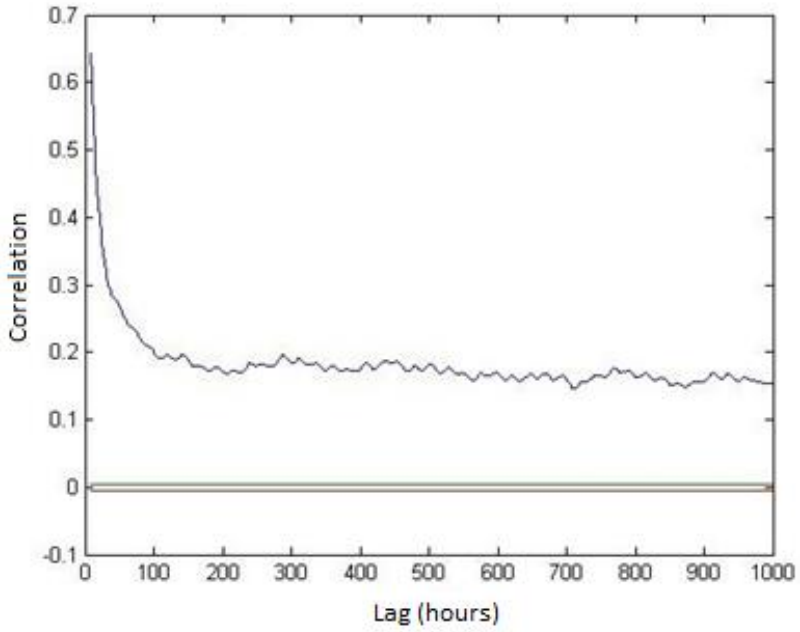

Fig. 5.14. The ACF for Lerwick, power transformed and diurnally detrended, lags 20-1000 hours

The figures show that the method has significantly reduced the diurnal seasonality, but has not eliminated it entirely. A next step was therefore to follow Daniel and Chen [128], as described in Chapter 4, and also divide the series by any diurnal profiles found for the standard deviation (s.d.). It was found that strong diurnal patterns for s.d. were generally present in the data, and that the profiles were close to the opposite of the profiles for mean, except that the peak daily difference occurred in the spring. The methodology involved was to:

- Calculate the s.d. for each hour of the day and each month of the year, i.e. create monthly s.d. profiles, then normalise them with the monthly mean values;
- 'Stretch' them into daily sd profiles using linear interpolation, as was done for the mean profiles; and
- Divide the wind speeds with these profile values.

Figure 5.15 below shows s.d. profiles for Peterhead Harbour (Zone 5), before normalisation, for two contrasting months. This is one of the stations where the effect is most prominent.

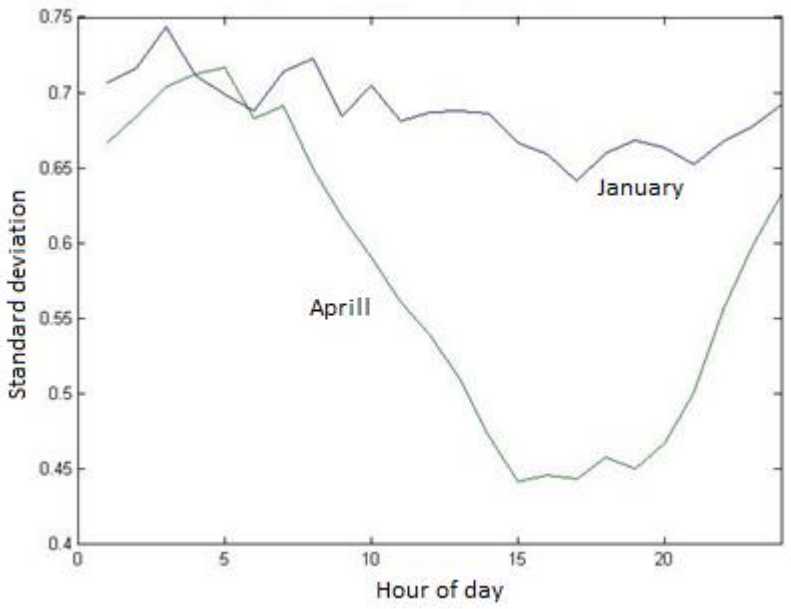

Fig. 5.15. The diurnal profile for standard deviation, for the transformed wind speeds at Peterhead Harbour.

The methodology was very effective in creating flat s.d. profiles, and eliminated most of the remaining diurnal seasonality that could be seen in ACF plots. The dataset was now ready to be modelled as a stationary Gaussian process, to a reasonably good approximation. Following the power transformation, low integer values were stretched further apart, while higher ones were squashed together, resulting in histograms that appear 'gappy' on the lhs. The removal of diurnal seasonality has served to counter this effect, resulting in much smoother Gaussian-like distributions. Returning to the example of Lerwick, the transformed and deseasonalised distribution is shown in Figure 5.16 below.

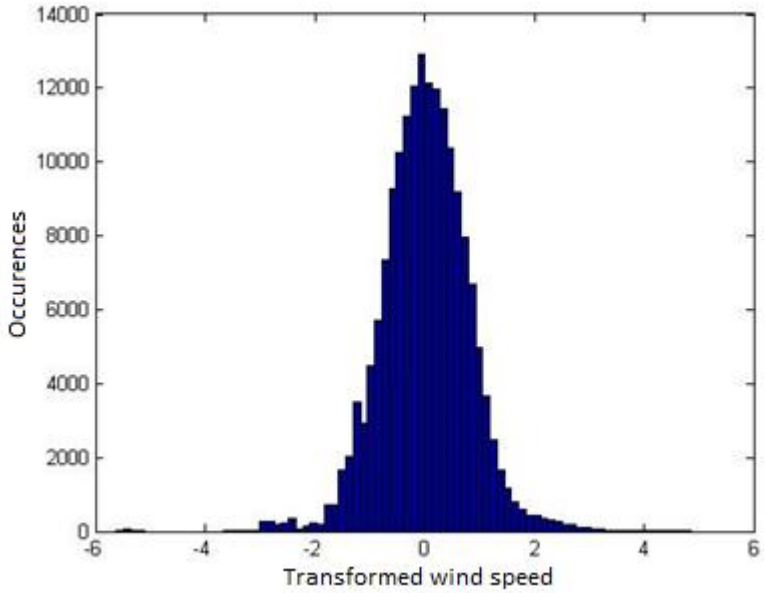

Fig. 5.16. Histogram for Lerwick, power transformed and deseasonalised.

5.4.3. The MVN and ARMA Methods for Gap Filling

Two methodologies were explored in order to fill-in the data gaps. The first, referred to here as the MVN method, is by far the simplest and makes no use of temporal correlations, only spatial ones. The methodology simply assumes that the transformed series is a MVN, with a known correlation matrix, and any missing values for each hour are taken to be their expectation values, conditional on the values at the zones for which valid data is present.

The ARMA method is considerably more complicated, and its major components are:

1) Fit high-order univariate AR models to each zone, and use these to create innovation series, i.e. series of forecast errors, for as many hours as possible given the data present;
2) Fit a VARMA model to the entire set of time series;
3) Move forward through the series, filling-in missing data as the VARMA model's 1-step-ahead forecasts. Model innovations form a MVN with a known covariance matrix and there should be added to the forecasts, importing as many innovation terms as are available from the univariate modelling, and use them to generate conditional expectation values for the unknown innovations.

The quality of the 'forecasts' will probably deteriorate as one progresses further into a gap, so:

4) Reverse the order of time for the series and generate a new set of hindcast errors using the univariate models (that are unchanged by time reversal);
5) Fit a new VARMA model to the time-reversed series (multivariate models are affected);
6) Re-fill in the gaps using the technique in step (3).

Since the temporal order has been reversed, the 'forecasts' made in this way should be most accurate towards the true end of each gap, so:

7) Re-fill the gaps again as a superposition of the two sets of values obtained above. The relative weighting of the two values must vary linearly with relative position within each gap, i.e. at the beginning the forward-moving values dominate, while at the end backwards-moving values dominate, and at the mid-point weights are equal.

It was initially assumed that for smaller gaps, the ARMA method will be most effective since it makes use of both temporal and spatial correlation structures. At the centre of longer gaps, however, the temporal correlations become less useful, and the relatively less important role of spatial information may put the method at a disadvantage. As an initial guess, gaps of 200 hours or longer were filled using the MVN method, shorter ones using ARMA. The longer gaps were filled first, in order to make the VARMA fitting process more accurate.

In order to test the accuracy of the MVN method, 10 sections of the dataset were found, 500 hours in length, with no missing data in any of the zones. The methodology was tested by assuming that one randomly chosen zone was empty for that section, filling it in, and then comparing the real and estimated values. The rms errors ranged from 0.40 to 1.08, with a mean value of 0.66. It was found that the error increases with the sample s.d., although the sample error/ s.d. ratio, which ranges from 0.58 to 0.80, is more favourable for higher s.d. samples. A value of 1 for this ratio would represent a method which is no more effective than sampling randomly from the $N(0,1)$ distribution. For nine out of the 10 samples, the s.d. of the predicted series was smaller than the true value, with ratios varying from 0.59 to 1.12, the average being 0.76.

When implementing the ARMA methodology, step (1) required the fitting of high order AR models – this was done using the Levinson-Durbin algorithm, described in Chapter 3, section 3.3.2. The two information criteria were used to select the optimal model order, p, finding that due to the very large sample size (175,298 hours) the AIC hadn't reached a

minimum value by $p = 10$, while the BIC became flat at $p = 7$. It was decided that $p = 10$ is the best choice, with no point increasing p further.

The fitting of the VARMA model for stage (2) was done using the Hannan-Rissanen procedure, also described in Chapter 4. The procedure required initial fitting of VAR models using Whittle's multivariate extension of the Levinson Durbin algorithm. It also required the fitting of models using ordinary least squares (OLS), which was achieved using *Matlab*'s Optimisation Toolbox. The optimisation problem was made easier by reducing it to 20 independent optimisation problems, i.e. minimisation of individual variances for the 20 one-step-ahead forecast error series. This is equivalent to minimising the trace of the multivariate error series' covariance matrix, rather than the determinant. Confidence that such a substitution could be made was gained from observing close similarities in the behaviour of the trace and determinant as model orders were increased, for AR models fitted using the multivariate Levinson-Durbin algorithm.

It was found that the best model order is VARMA(3,1), since this is the model with the lowest BIC value. Residual ACFs were plotted and found to be uncorrelated to an acceptable level. It was found that the model is stationary, but only just – some roots of the determinant of the AR polynomial are very close to unity. It was similarly found that the model is invertible, with solutions more comfortably clear of the unit circle. For the model required for step (6) of the ARMA methodology, it was assumed that the optimum model order has not changed, and the model was fitted in the same way – Levinson-Durbin and OLS using *Matlab*.

It was found that using the ARMA method to fill-in the 500 hour long gaps described above provided surprisingly similar results to using the MVN method. Indeed, the zonal ratios of ARMA rms errors to MVN rms errors only ranged from 0.9288 – 1.1055 for the 10 samples. Even more surprisingly, the ratios remained very similar when the length of the samples was reduced to 100 hours. Figure 5.17 shows a comparison of the true values of the (transformed) series for Wick and values filled-in by the ARMA method. These are the first 20 hours of a 500 hour gap, so the superposition of the backwards and forwards series is very heavily biased towards the forwards model. The figure shows that the accuracy seems to be a function of how quickly the real series changes, rather than distance from the beginning of the gap.

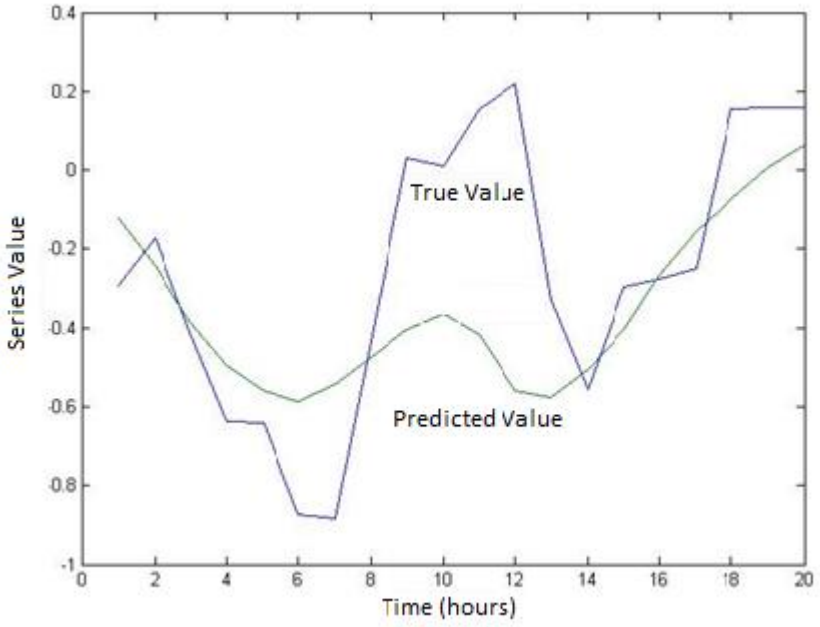

Fig. 5.17. Example of real vs. ARMA method filled in values, Wick.

This is confirmed by Figure 5.18, which shows the r.m.s error as a function of distance from the start of a gap, averaged over the 10 samples, for Wick.

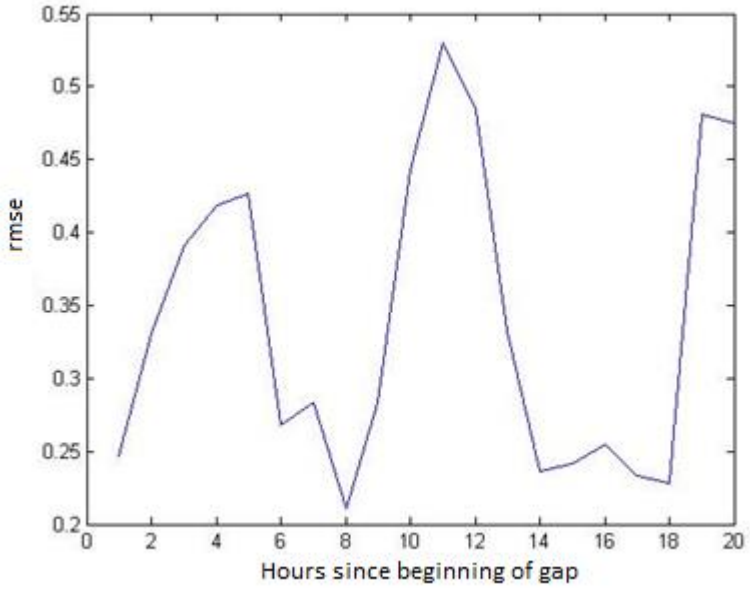

Fig. 5.18. RMS error of the ARMA filling method vs. distance into gap, for Wick.

It is clear from Figure 5.17 that this method also reduces the s.d. compared to the true series. The extent of this effect is very similar to that of the MVN technique for some samples, but is somewhat worse for others. Since the effectiveness of the two methodologies are practically indistinguishable in other respects, the conclusion to be drawn is that the MVN method is preferable for gaps of all lengths, and therefore all gaps in the series were filled in this way.

Having filled-in every missing datum, the detrending, standardisation and power transform were reversed. The research project has therefore produced a 100% complete 20 year historical time series representing the wind speed field over Great Britain, prepared in a form which makes calculations in *Matlab* easy. While this dataset may be used by present and future researchers at the University of Bath, it cannot be shared in an entirely open way due to the data sharing agreement with the BADC. There is however no restriction on sharing the results of analysis conducted on the series, and this is precisely what is done in Chapter 6.

5.5 Chapter Summary

This chapter described the process of wind speed data acquisition for this project, from the BADC, and subsequent initial analysis and processing. The data was found to have many quality related issues initially, but these were resolved with the exception of excessive zero speed recording due to anemometer friction and freezing etc. Several attempts to separate true and spurious zero wind speed recordings were unsuccessful.

Locations were found for each zone where a very high percentage of valid data was present throughout a 20 year period; the data were also free of problems such as single long gaps. Careful examination showed that while the exact locations of the recording equipment are not entirely ideal in some cases, there do not seem to be any serious problems. Two methods were explored for filling in the missing data – one simple and the other much more complicated, finding that the simple method is slightly better. Having filled in the missing data, an entirely complete 20 year historical record representative of the GB wind field is now available for academic use.

The decision was made to power-transform then centre the series before model fitting, rather than taking logs. When reversing the process during simulation, negative values following the re-introduction of the mean will simply be set as 0 before power transformation. In order to establish the best common value for power transformation, certain zones were identified as being more important in the sense that they will almost certainly contain much more wind capacity than others in the future.

It was found that during summer months, many zones display quite prominent diurnal patterns in mean. Even more pronounced patterns were found in standard deviation, strongest in spring. It was shown that these patterns may be successfully modelled as deterministic in nature and removed accordingly. This was in fact necessary before filling-in missing data, to improve accuracy. The removal of such patterns had the added benefit of ‘breaking-up’ integer wind speed values, such that transformed wind speed distributions appear much more natural.

Chapter 6

Analysis of the Historical Dataset

This chapter reports on the results of initial data analysis on the chosen 20 year samples, confirming some of the main features of the resource discussed in Chapter 2, and enhancing understanding of others. Many results confirm the validity of modelling choices tentatively made in Chapter 3, but the chapter also highlights the complexity of patterns of resource availability on various time-scales and the extent of the modelling challenge.

This chapter also discusses attempts to simplify and interpret those complex patterns through principle component analysis (PCA) and clustering. It then moves on to examine the chosen period in terms of the reduced GWLs discussed in Chapter 2 and explore the extent to which the most likely reduced GWL states for a given day can be implied from the principle components' values. This leads to discussion of the feasibility of a methodology to be developed in the future that would enable synthetic wind speed field values to be correlated to synthetic electricity demand values with a degree of accuracy beyond consideration of the day of week and day of year.

6.1. Statistical Properties of the Series

6.1.1. Station Wind Speeds and their Cross-Correlations

Table 6.1 below presents summary statistics for the distributions of wind speeds at the chosen set of stations, as represented by the filled-in raw data. The values were calculated for the entire 20 year period and are therefore treated as asymptotic values. Examining the mean values reveals that sites in the north and west are generally the windiest, as expected, although local factors such as proximity to the sea and elevation are also factors. The means vary from around 7 knots at Northolt (greater London) to nearly 15 knots at Lerwick in the Shetland Islands. Standard deviation shows a similar level of variability between locations, and a strong positive correlation exists between mean and standard deviation. Reflecting this,

Northolt has the smallest standard deviation, under 5 knots, while it is greater than 8 knots at Lerwick.

Zone	Name	Mean Wind Speed (Knots)	Standard Deviation (Knots)	Skewness Coeficient
1	Lerwick	14.6569	8.1283	0.8125
2	Stornoway	11.5057	6.9798	0.8996
3	Wick	11.0503	6.1893	0.9084
4	Tiree	14.143	7.3912	1.0427
5	Peterhead Harbour	11.3689	6.4165	0.8938
6	Dunstaffnage	8.0741	5.0653	0.6406
7	Machrihanish	12.0964	6.9954	0.9650
8	Salsburgh	12.5884	6.668	0.9839
9	West Freugh	10.1074	6.3531	0.7667
10	Carlisle	7.6325	5.7041	0.7683
11	Valley	12.0575	7.6238	0.9481
12	Leeming	7.9235	5.4732	0.5560
13	Waddington	8.9378	5.0465	0.6903
14	Aberporth	13.2793	7.5322	0.9545
15	Shawbury	8.1075	5.0276	0.6060
16	Marham	8.9538	5.3313	0.7536
17	Cambourne	10.7677	6.156	0.6155
18	Solent	11.7928	7.2166	0.8080
19	Northolt	6.991	4.6586	0.7122
20	Manston	9.4373	5.1695	0.7735

Table 6.1. The mean winds speeds, standard deviations and skewness coefficients for the chosen stations.

Table 6.1 also presents (Pearson's) skewness coefficients for the distributions, also referred to here as simply 'skewness'. The coefficients are given by $E[(X - \mu)^3/\sigma^3]$, and are a measure of the asymmetry of a distribution. Unimodal distributuions where the left tail is fatter or longer have negative skewness, while the opposite is true if the right tail is fatter or longer. For multimodal distributions - as is the case, to some extent, for several zones – the meaning of a skewness vlaue is rather less clear. Since normal distributions are symmetrical, skewness can be used as a measure of deviation from normality – although not an entirely reliable one, since asymmetrical distributions can have zero-valued skewness coefficients.

Table 6.1 shows that all locations have positively skewed distributions, as might be expected. The smallest skewness value is 0.5560, while 9 stations have values between 0.6 and 0.8, a further 9 are between 0.8 and 1, and the largest value is 1.0427.

The intention of the power transformation was to make the distributions at each location approximately normal, and thus should bring skewness values close to zero. This was not possible in practice, partially due to the compromised nature of the transformation coefficient, but more so because of the smaller peaks at low or zero wind speed, which diurnal detrending can only spread out, rather than eliminate. After transformation and diurnal detrending, 8 zones had skewness coefficients between +0.1 and -0.2, 5 zones were between -0.3 and -0.5, and 7 between -0.5 and -1. Most zones' skewness therefore became closer to zero, some significantly so. The worst (i.e. furthest from zero) is zone 19, which is fortunate in the sense that it has very little capacity according to the scenarios discussed in Chapter 4 and the RenewableUK map. Figure 6.1 shows an example of a zone left with negative skew due to the small peak to the left of the centred main peak.

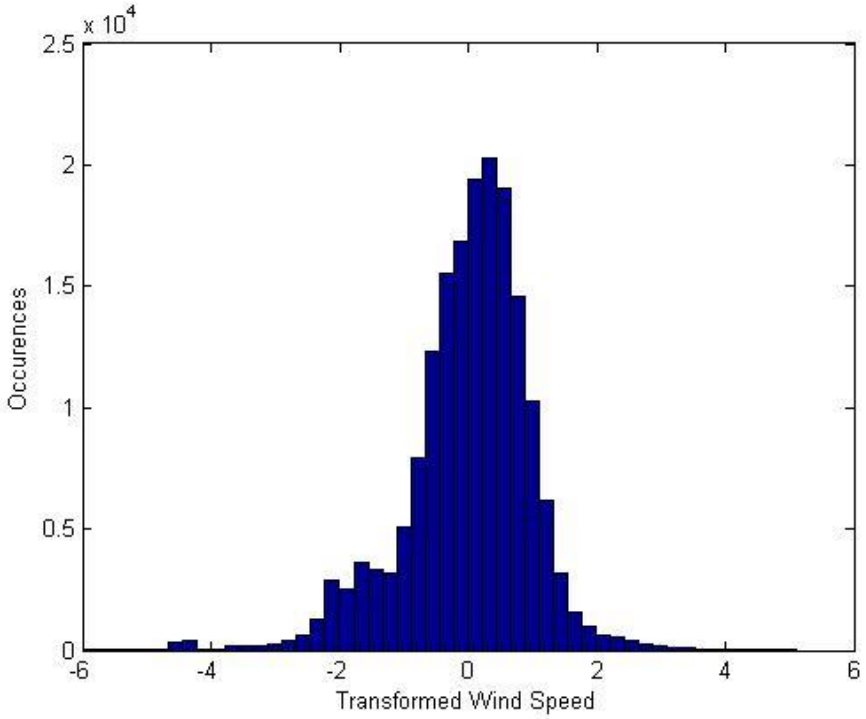

Fig. 6.1. Histogram for the transformed series, Northolt.

Table 6.2 shows the long-run linear correlation coefficients for hourly wind speeds at each pair of sites. The table shows that sites are generally highly correlated, whith no negative values. The cross-correlations vary between a maximum value of 0.8 for Zones 7 and 9, that are very close together, and a minimum of 0.08 for Zones 2 and 20, which are diagonally opposite to each other across the country. Interestingly, this value is smaller than the correlation of 0.13 between Zones 1 and 17, easily the pair separated by the greatest distance. The mean correlation is 0.49, and 46% are greater than 0.5. Summing all cross-correlations for each zone shows that Zone 1 has the smallest total at 6.03, reflecting the location's geographical isolation, while the largest is 10.33 for Zone 9, located half way down the country and close to several other stations. There are no standout locations with an unexplained smaller total – a strong indication that there are no 'dud' sites due to orographic forcing or sheltering.

Zone	1	2	3	4	5	6	7	8	9	10
1	1	0.51	0.64	0.42	0.55	0.4	0.35	0.43	0.34	0.37
2		1	0.58	0.65	0.53	0.5	0.46	0.52	0.45	0.4
3			1	0.55	0.68	0.6	0.51	0.5	0.49	0.45
4				1	0.59	0.7	0.72	0.59	0.67	0.49
5					1	0.5	0.53	0.57	0.56	0.45
6						1	0.72	0.67	0.67	0.62
7							1	0.65	0.8	0.6
8								1	0.65	0.76
9									1	0.66
10										1

Zone	11	12	13	14	15	16	17	18	19	20
11	1	0.61	0.61	0.71	0.66	0.54	0.49	0.52	0.44	0.33
12		1	0.72	0.53	0.69	0.66	0.4	0.5	0.55	0.43
13			1	0.58	0.74	0.81	0.49	0.63	0.69	0.57
14				1	0.65	0.56	0.68	0.59	0.51	0.43
15					1	0.72	0.54	0.64	0.68	0.52
16						1	0.51	0.7	0.79	0.71
17							1	0.59	0.54	0.44
18								1	0.76	0.65
19									1	0.75
20										1

Table 6.2. The correlation coefficients for each pair of stations.
Dark grey: correl. > 0.5, light gray: 0.25 < correl. =< 0.5.

6.1.2. Time Series Plots

The most basic form of analysis carried out on the series was to simply to plot sections of it. An example of the 'raw' series was given in Chapter 5, and will not be repeated here. In order to explore very low frequency variability, Figure 6.2 below shows a plot of the 'raw' series reduced to annual averages, for 3 zones. The plot reveals quite significant variability on this time-scale. Patterns appear stochastic, but with considerable autocorrelation, and much of the oscillations have periods on the scale of decades. The gradual increase in wind speed at

Stornoway from 1993 until the end of the sample is interpreted as a very low frequency oscillation, rather than predictable climate trend. This is based on the fact that the increase is not seen at the other locations, and the exceptionally poor wind energy yields reported for the entire UK in 2010 [75]. It was found that cross-correlations at this aggregation level are generally small and positive, with some slightly negative.

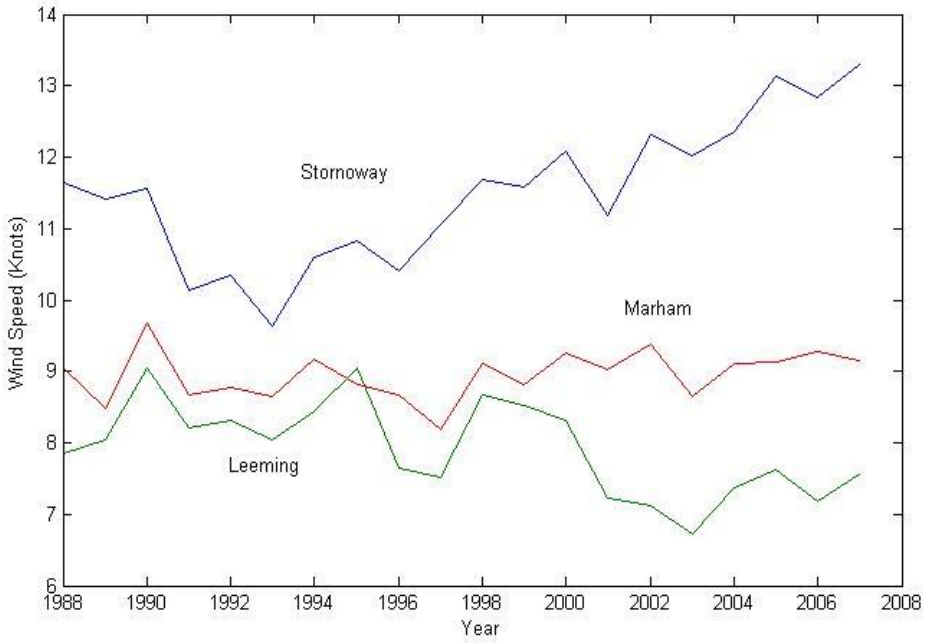

Fig. 6.2. Plots of the 'raw' series, annually averaged

Figure 6.3 shows the series of monthly averages for the raw series at Lerwick (with the 20 year mean removed), for two very different years separated by a decade. Winter months were the windiest for both years, but the difference between winter and summer is much greater for the windy year, and the patterns clearly appear more stochastic than deterministic in nature.

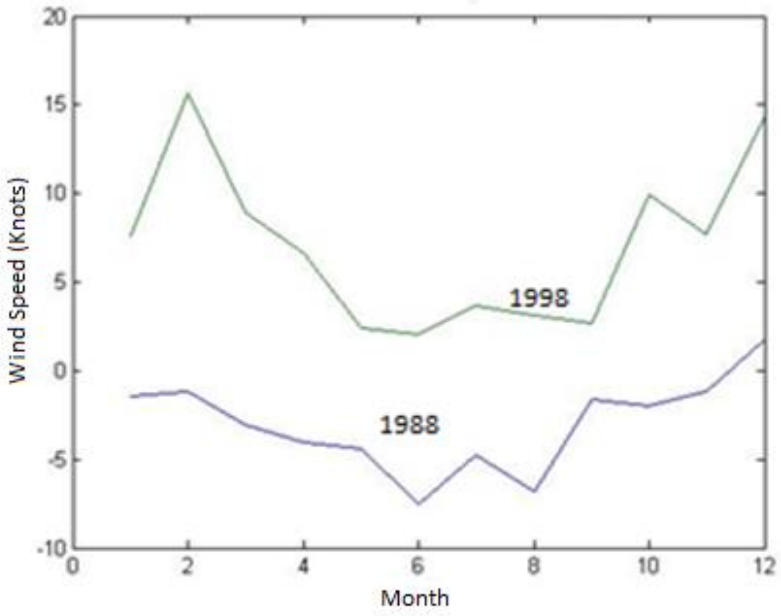

Fig. 6.3. Monthly averages for contrasting years, centred raw series, Lerwick.

The same can be said for variance calculated on a monthly basis, as shown in Figure 6.4 below for the raw series, again at Lerwick, with values divided by their 20 year mean. The figure shows that the monthly variances range by a factor of 10 at least. The power transformation had the effect of partially stabilising variance such that the ten-fold difference seen between a pair in the figure was reduced to a factor of four, for example.

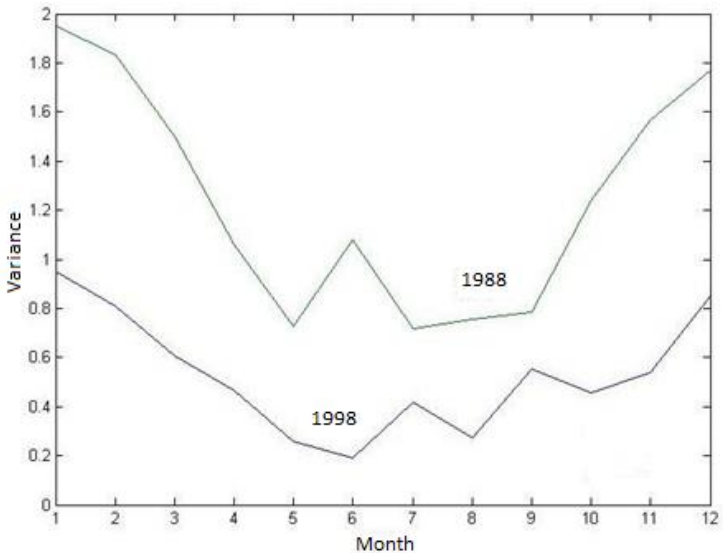

Fig. 6.4. Monthly variances for contrasting years, raw series, Lerwick.

Figure 6.5 below shows a monthly mean series for wind speeds and concurrent standard deviation, in this case covering a period of 3 years, for Valley. The clear irregularity of the annual periodicities seen here confirms their stochastic nature. There is clearly a strong positive correlation between the series during this period, and several other plots generated for different zones showed that this plot is representative.

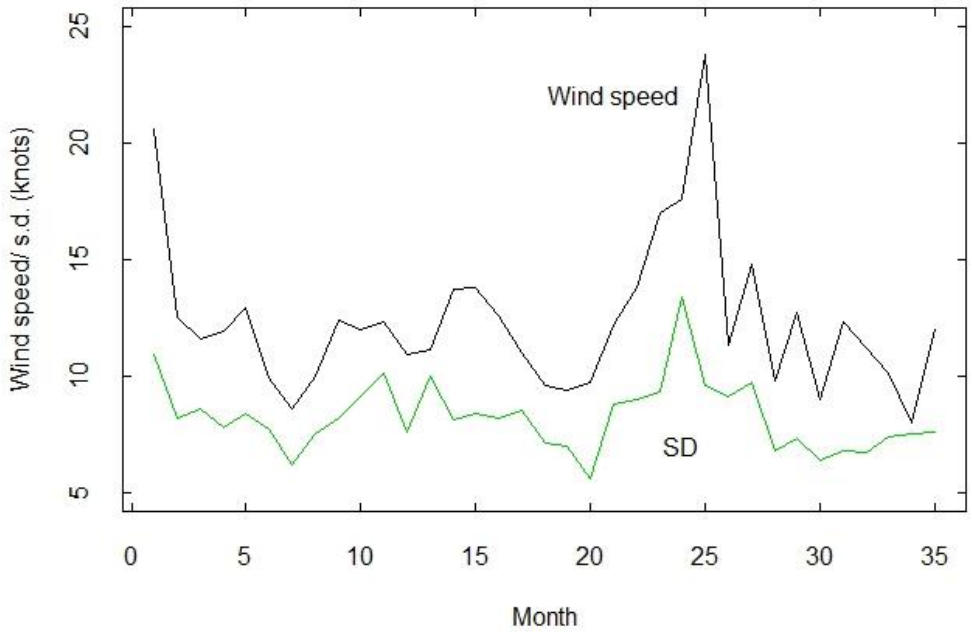

Fig. 6.5. Monthly means and standard deviation over a 3 year period, raw series, Valley.

6.1.3 Differences between Individual Months

The plots above indicate that individual months can be very different from each other. One aspect of interest is the relationship between monthly means and variances. This is explored further via a scatter plot of monthly values for a transformed and standardised series in Figure 6.6 below, which uses the entire sample for Camborne. The x-coordinated show that monthly means are roughly Gaussian distributed, but variances are not – a Gamma distribution would appear to be more appropriate. The transformation appears to have weakened the correlation between mean and variance – the correlation coefficient is 0.23.

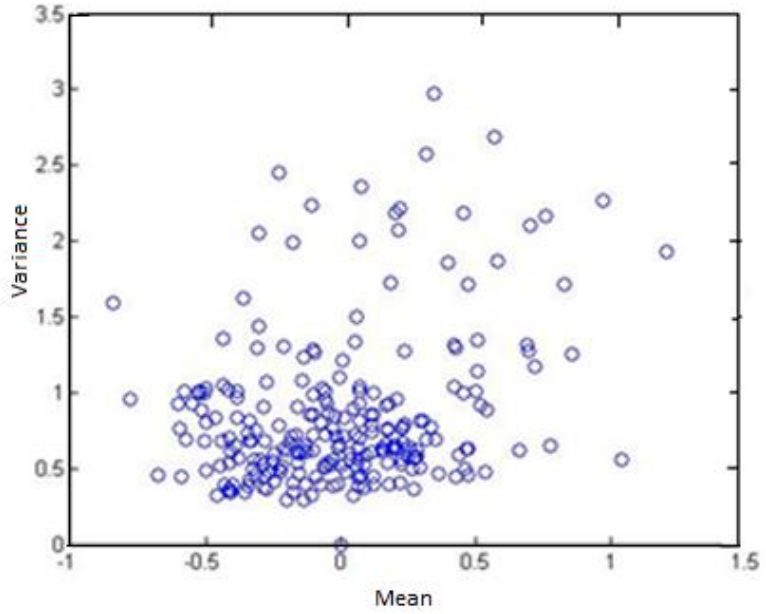

Fig. 6.6. Scatter plot showing the joint distribution of monthly means and monthly variances, across the power transformed series, Camborne (Z17)

Plotting histograms for individual months revealed a wide variety of distributions, many highly irregular in shape, despite relatively large sample sizes of $n = 672$ to 744. Figure 6.7 below shows a somewhat extreme example month for the transformed and standardised series at Tiree.

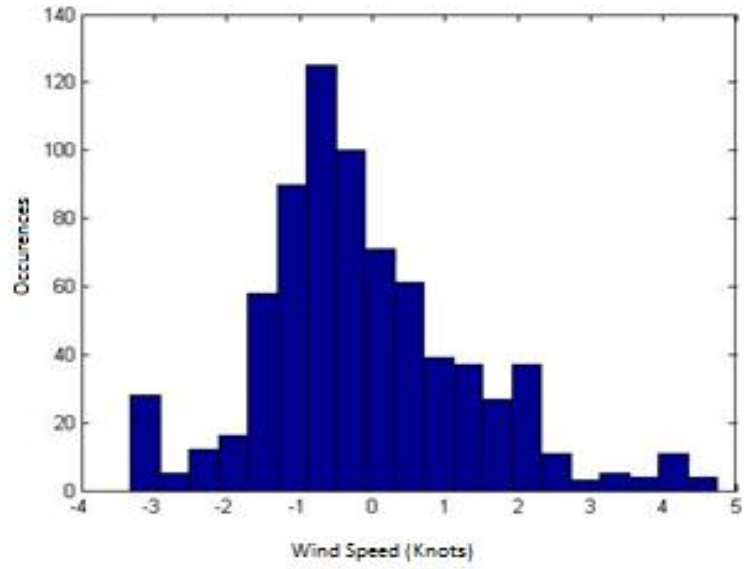

Fig. 6.7. Example histogram for a month, transformed series, Tiree.

The variability of wind distribution shapes from month to month is reflected in monthly skewness coefficient values. Their variability is demonstrated below in Figure 6.8, via a scatter plot, again using every month for Camborne, transformed. The skewness coefficient for the 20 year sample is -0.3689, but for individual months values vary from about +1 to -1.8. The plot shows skewness vs. mean wind speed, in order to explore their relationship. The plot suggests that they do not have a strong relationship, and indeed their linear correlation is +0.17.

There does not seem to be a way of directly ensuring that such subtleties in the series' behaviour are replicated in the synthetic data. The extent to which such dynamics are implicitly captured by model structures, and reproduced during simulation, is an important aspect of model validation.

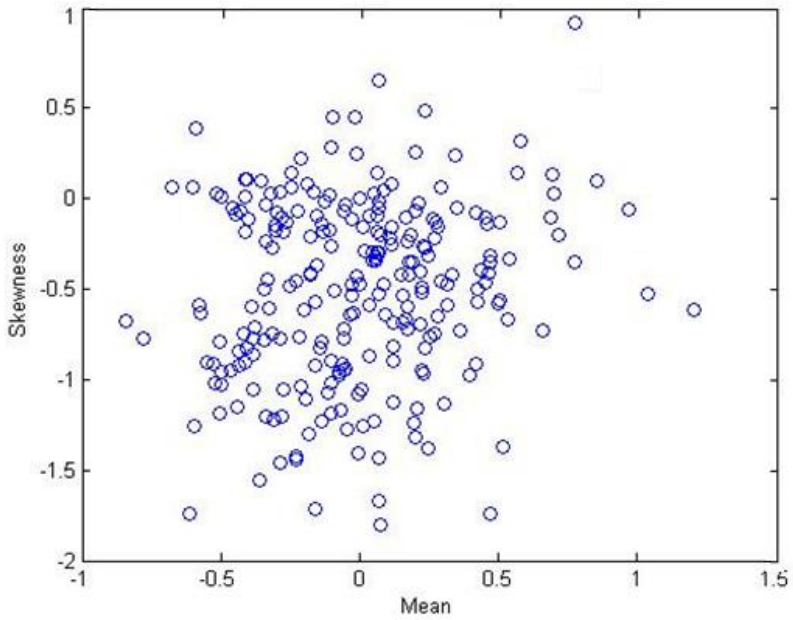

Fig. 6.8. Scatter plot showing the joint distribution of monthly means and skewness coefficients across the entire transformed series, Camborne (Z17).

6.1.4. Correlograms

This section reports on the plotting of auto-correlograms and cross-correlograms for the series. The first to be discussed are the most simple: short-lag auto-correlograms, as shown in Figure 6.9 below for Lerwick, raw series. A rapid decay in the sample autocorrelation function (ACF) is clearly observed up to a lag of about 36 hours, although by a lag of about 80 hours it is clear that decay to zero will be very slow. Diurnal seasonality appears quite subtle here due to the scales on the axes. The white noise boundaries discussed in Chapter 3 have been included in the plot, and represent very small correlations, due to the large sample size.

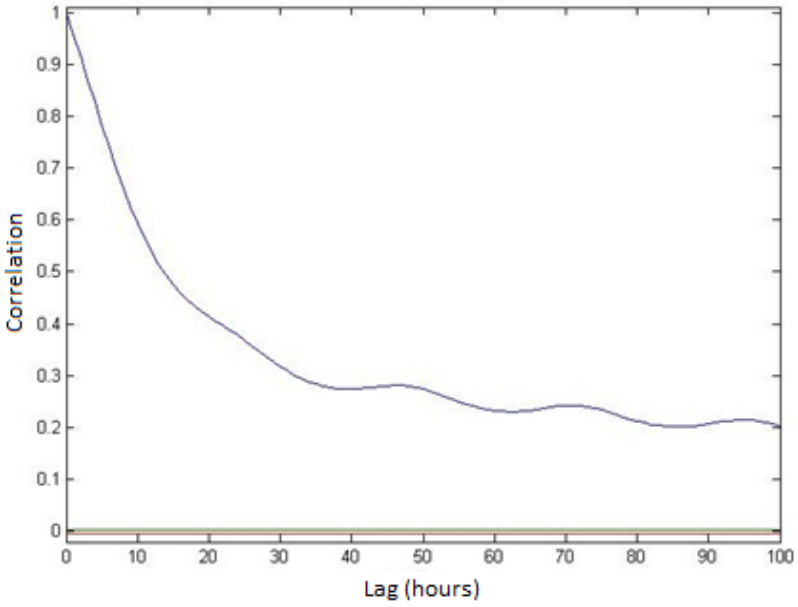

Fig. 6.9. ACF for lags 0 – 100 hours, raw series, Lerwick.

Figure 6.10 shows the cross-correlogram for Lerwick and Leeming, over the same range of lags, raw series. One function has Lerwick leading, the other following – the latter has larger values for positive lags, presumably since the direction from Leeming to Lerwick (almost exactly north) is partially aligned with the average direction of fronts across GB. Both functions are very similar to the ACF for Lerwick, except scaled down by roughly a factor of 4 and with relatively larger diurnal seasonality.

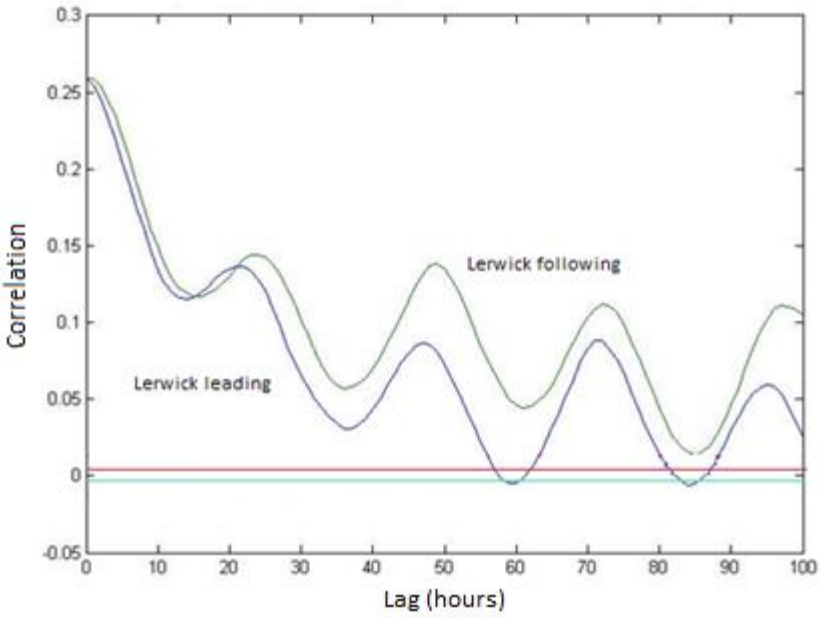

Fig. 6.10. Cross-correlation functions for Lerwick and Leeming, lags 0 – 100 hours.

Figure 6.11 shows a set of correlation functions for the transformed series, up to 24 hours lag. One is the ACF for Lerwick, the others cross-correlation functions (CCF's) of Lerwick (Z1) with Stornoway Airport (Z2), Wick (Z3), Tiree (Z4) and Peterhead Harbour (Z5), all with Lerwick lagging. The figure shows that cross-correlations have decreased as a result of transformation, and have their maximum for non-zero lag. At zero lag, CCF's decrease with distance, as expected, but by a lag of 8 hours their order has changed. The CCF peaks occur at greater lags for larger distances, again as might be expected. This highlights the possible advantage of choosing a higher order model in the multivariate case.

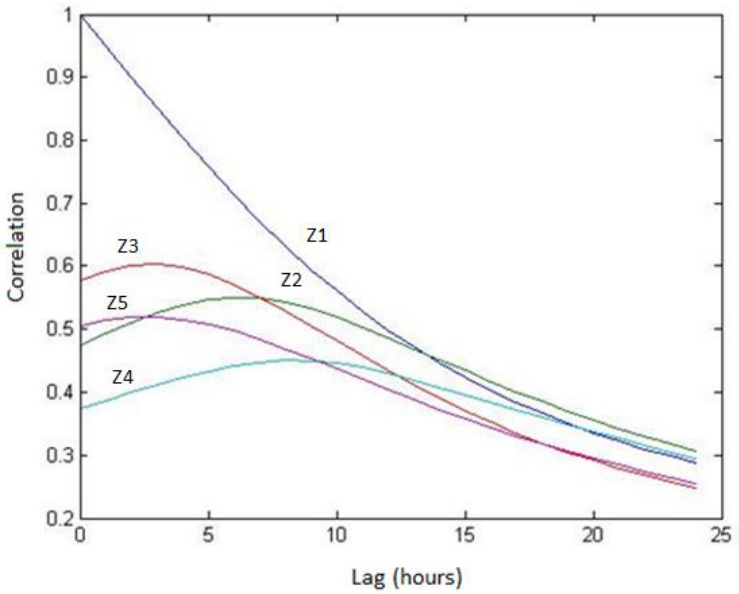

Fig. 6.11. ACF for Zone 1 and CCF's for Zone 1 with Zones 2, 3, 4 & 5, transformed series, lags 0 – 24 h.

Partial-correlograms were also plotted for the transformed series, a typical example of which is shown below in Figure 6.12, for Wick. It shows that while the function decreases to a very small value rapidly, it does not fall within the white noise boundary even for a lag of 10 hours – consistent with the discovery discussed in Chapter 5 that the univariate model AIC did not reach a minimum up to a model order of $p = 10$.

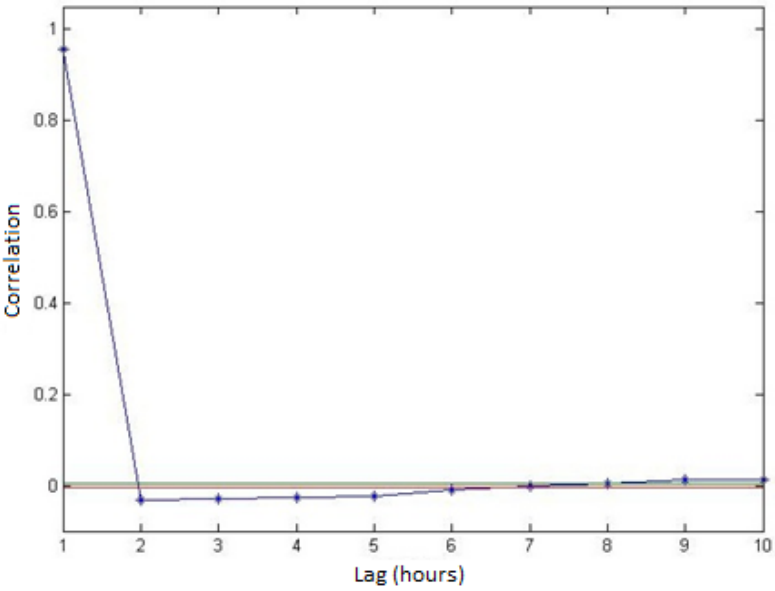

Fig. 6.12. Partial-ACF for Wick, transformed series, lags 1 – 10 hours.

Auto-correlograms were plotted up to very long lags – longer than in the two papers exploring long memory in wind time series discussed in Chapter 4, [88] and [90]. An example is shown in Figure 6.13 below for Lerwick, transformed series, lags 100 – 12,000 (~1.5 years). Clearly the annual seasonality dominates, but it is interesting that the ACF only just crosses over into negative values for lags of about 6 months, and certainly the oscillation envelope is much wider than the noise boundaries. If the seasonality were purely deterministic, i.e. essentially not changing in nature from year to year, then roughly half of correlations should be negative. The very slow rate of decay for the oscillation envelope stongly suggests that the series is 1st order non-stationary, and the inclusion of long memory is sensible, as the literature suggests.

Long-range auto-correlograms for raw series are very similar, except that the curves appear thicker due to the diurnal seasonality.

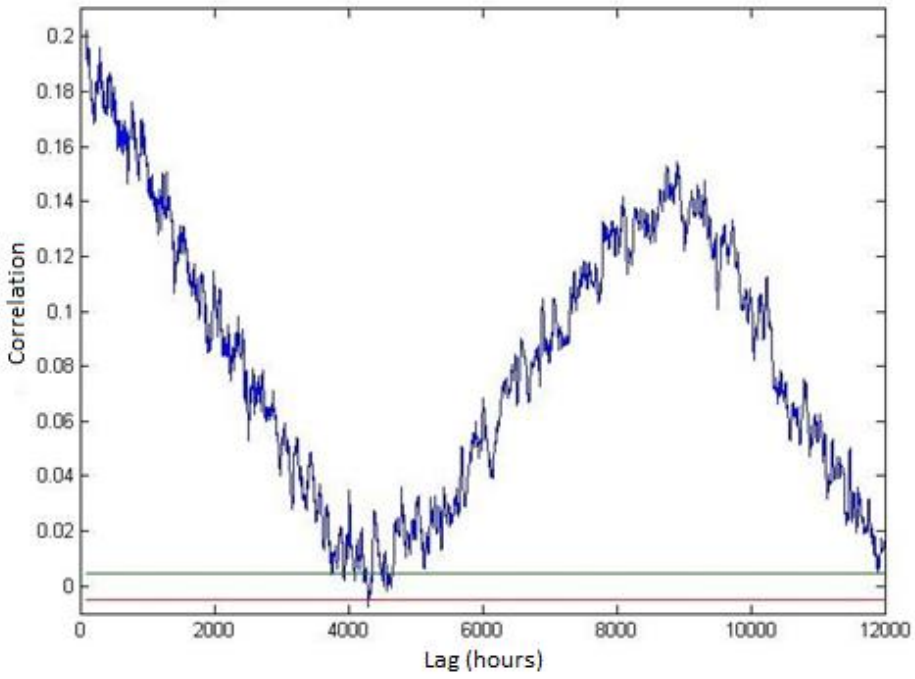

Fig. 6.13. ACF for Lerwick, transformed series, lags 100 – 12,000 hours.

6.1.5. Periodograms and Long Memory

Periodograms were plotted for each zone's transformed series, using equation 3.13 from Chapter 3. For some zones, a single unambiguous pole at the annual frequency suggests very strongly that the series should be modelled as a Gagenbauer process, as suggested by Bouette et al. [90]. An example is Valley, Zone 11, as shown in Figure 6.14., with frequency on a log scale.

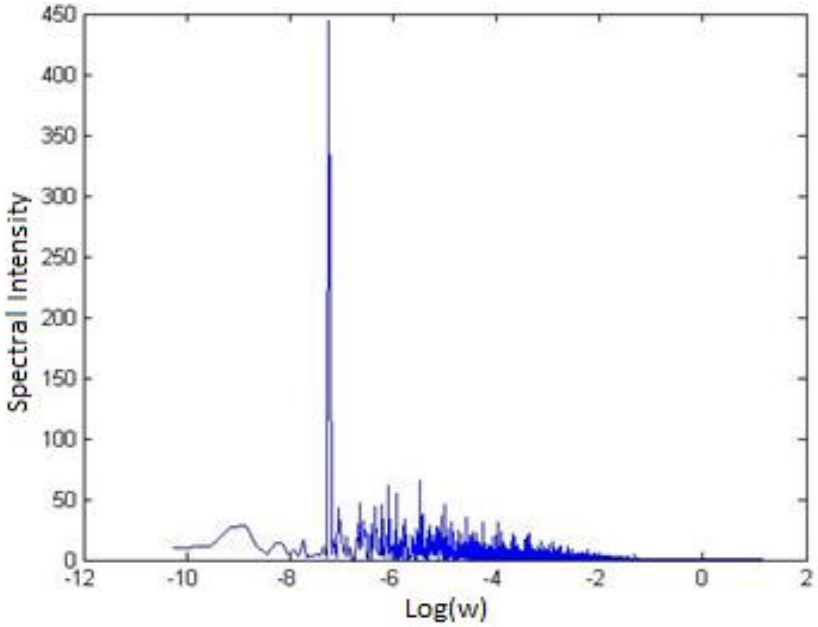

Fig. 6.14. Periodogram for Valley (zone 11), transformed series, log-linear axes.

The situation is not so clear in other zones, unfortunately. The periodogram for Lerwick, for example, is shown in Figure 6.15, with log scales for both axes in this case. As in several other zones' periodograms, the spectral intensity at the annual frequency is not so obviously greater than the low frequency extreme. The same plot with a linear y-axis shows that the annual frequency remains the maximum, but the difference is reduced.

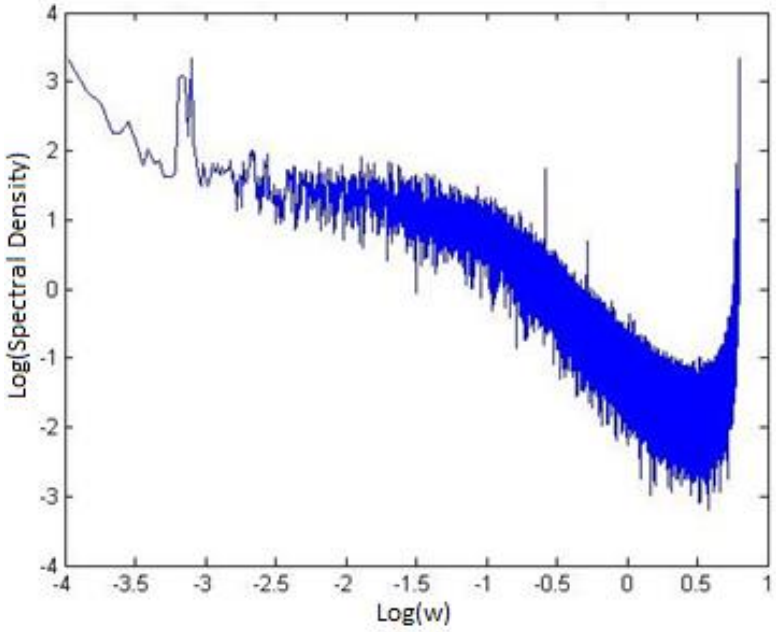

Fig. 6.15. Periodogram for Lerwick (zone 1), transformed series, log-log axes.

The strength of the spike for Valley is interesting since, as discussed in Chapter 2 (section 2.4.2), a study by Palutikof et al. [80] found that this station's wind speeds displayed significant correlation with an 'anticyclonicity' index for the greatest number of months. This characteristic may be regarded as reflecting how non-stationary the series is, which in turn may be interpreted as the strength of long memory present. The mathematics of Gagenbauer processes suggest that for finite samples, the height of the spikes reflect the value of the differencing parameter, i.e. the extent of long memory. The fact that the locations at which we would expect to find the strongest long memory, on the basis of [80], are indeed those with the tallest spikes provides a (rather week) validation of the idea that the Gagenbauer process assumption is appropriate.

For a process exhibiting regular (i.e. non-seasonal) long memory, the spectral intensity should be a straight line for several of the lowest octaves in the log-log periodogram plot – and this is roughly the case for Figure 6.15. Considerable doubt therefore remains on whether modelling the set of wind time series as a multivariate single-frequency Gagenbauer process is the best way of capturing all aspects of their behaviour.

Another complexity is that the long memory parameter itself may be seasonal in nature, stochastically or deterministically – an aspect that obviously cannot be captured by a single process with fixed coefficients. The seasonal nature of the parameter value is evident

from examination of the variance of sample means as a function of sample size. Plotted on a log-log scale, a regular long memory process would display a decrease in variance with sample size, with the relationship following a straight line. The gradient of this line gives a 1st approximation of the long memory parameter, as discussed in section 2.5.1. Examination of monthly mean time series plots, discussed in section 6.2 above, showed that winter months are more variable than summer ones.

To connect this to the long memory parameter, samples may be constructed, centred about the 1st of January and gradually increasing in size from 1 to 12 months. The log variance of these sample means (over the 20 years in the sample) can then be plotted versus the log of sample size. The procedure may then be repeated for samples centred on the 1st of July. If there is a large difference between the gradients of these two lines, then the long memory parameter is truly seasonal. This is indeed the case for the GB wind speed data, as shown for Machrihanish, Zone 7, in Figure 6.16. This will simply have to be accepted as a shortcoming of the modelling methodology, as the total number of parameters involved in fitting seasonal parameters is unacceptably high.

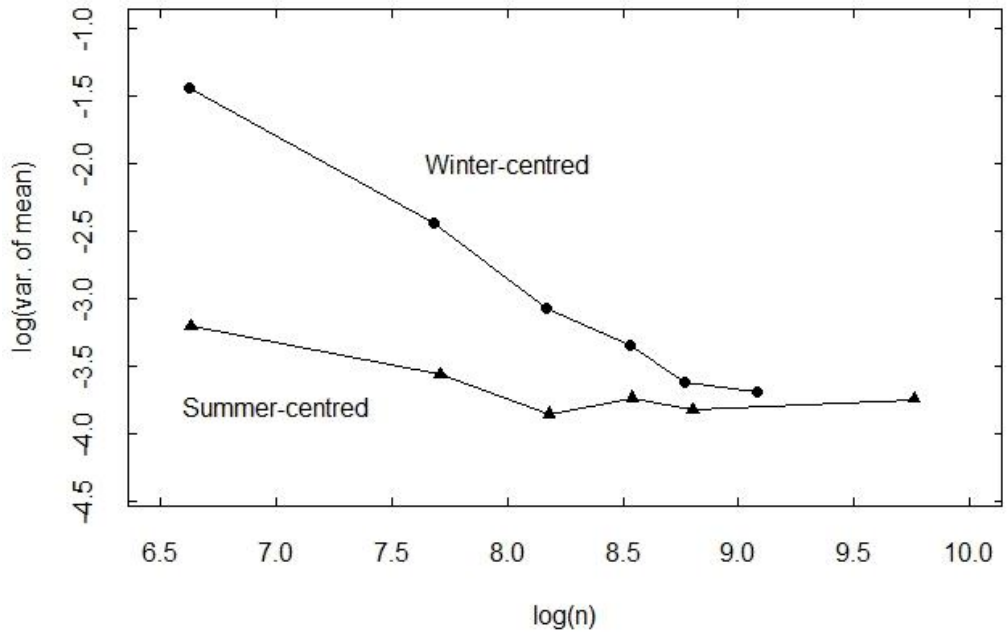

Fig. 6.16. Log variance of sample means vs. log sample size, for summer and winter centred samples, transformed series, Machrihanish (zone 7).

6.2. Principle Component Analysis

The completed, transformed 20-year series was further transformed into principle component form, following the method described in Chapter 3. It was seen that the 1st component dominates, accounting for 43.71% of the total variance, with contributions falling quite rapidly to 14.88% for the 2nd component and 5.95% for the 3rd. In order to account for more than 75% of the total variance, only the first 6 components are needed, while components 1-10 collectively account for 85.16% of it.

Table 6.3 below provides 3 example covariance matrix eigenvectors, corresponding to the 1st, 3rd and last components, and rounded to the nearest single decimal point for ease of comparison. The elements of the principal component eigenvector are all very similar, demonstrating that the principal mode of variability is roughly the spatially averaged wind speed across the country. The other two involve regional oscillations, with some zones almost irrelevant.

It was stated in Chapter 3 that the multivariate modelling task probably could not be simplified by working with a reduced number of principle components, since their dynamics are very different, with the first components' spectra dominated by lower frequencies, and the opposite for the last components. Figures 6.17 and 6.18 below show concurrent 200 hour time series for the 1st and 20th components, and the difference between them is indeed striking, apparently confirming the hypothesis.

Zone	Eigenvectors		
	1	2	3
1	0.2	0.5	0
2	0.2	0.2	0.1
3	0.2	0.4	-0.1
4	0.2	0	-0.2
5	0.2	0.3	0.2
6	0.2	-0.1	0.1
7	0.2	-0.2	0.4
8	0.2	-0.2	-0.5
9	0.3	-0.3	-0.4
10	0.2	-0.2	0.4
11	0.3	-0.2	0.3
12	0.2	0	0
13	0.2	0	0.1
14	0.3	-0.2	-0.2
15	0.3	0	0
16	0.2	0.2	-0.3
17	0.2	0.5	0
18	0.2	0.2	0.1
19	0.2	0.4	-0.1
20	0.2	0	-0.2

Table 6.3. Example eigenvectors of the historical series' covariance matrix, rounded for ease of comparison.

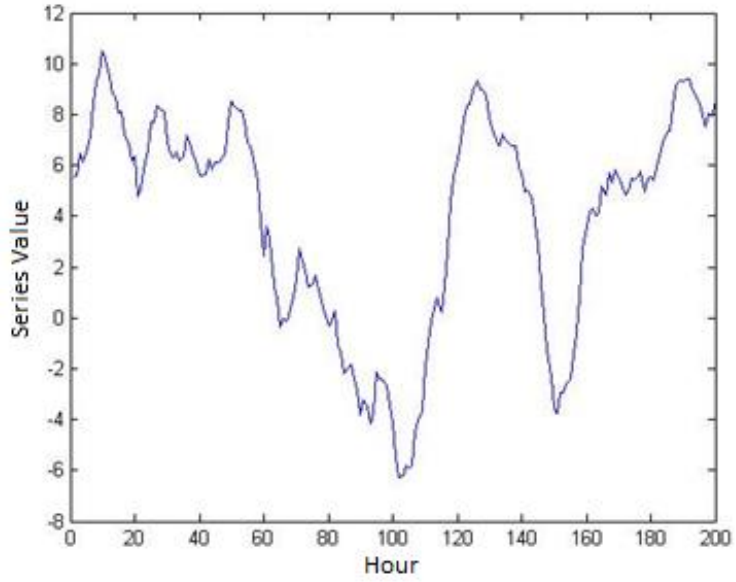

Figure 6.17. Time series segment for the 1st principle component.

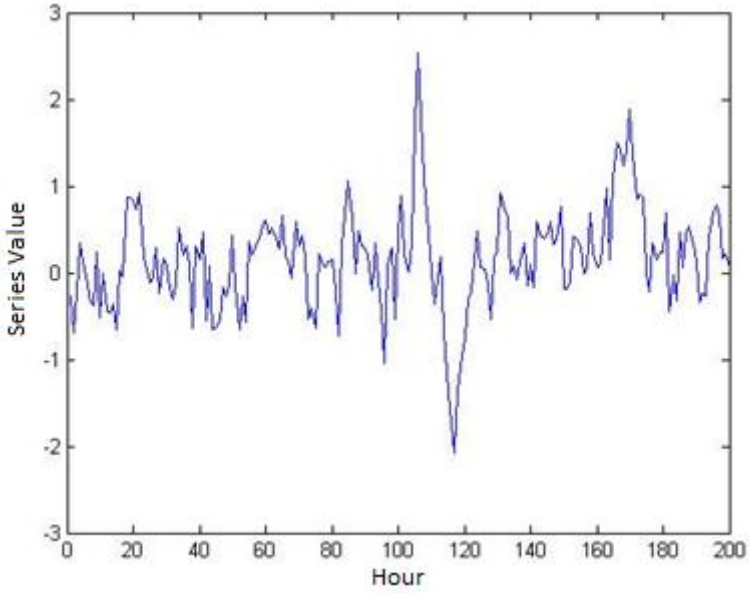

Figure 6.18. Time series segment for the 20th principle component.

As described in section 6.1.3, the distributions of most zones' power transformed series have some negative skewness due to the smaller peak to the left of the main one, which have merely been spread out by diurnal detrending. It is interesting that each principal component has only one peak, and therefore their distribution skews are generally closer to zero. Their skews are not exactly zero however, and this is usually due to a few very extreme values. The 1st principal component has some positive skewness, and looks slightly 'Weibull-like', reflecting the necessarily compromised choice of the power transformation coefficient.

6.3. Reduced GWL Circulations During the Sample Period

6.3.1. Incidence Rates for the Circulation Types during the Sample Period

This section makes use of data provided by Dr David Brayshaw from the University of Reading's Department of Meteorology. Dr Brayshaw and colleague Dr Giacomo Masato have used the methodology of James in [70] to recreate his series of objective GWL atmospheric circulation type classifications, as reported in [77]. The data made available was a daily series of GWL types for the entire 20 year fitting period 1988-2007. The series was then converted to have only the reduced number of GWL types described in [77] and Chapter 2. From this, time series consisting of the % of summer/ winter days attributed to each circulation type during each year of the period were created, and the winter series are shown in figures 6.19 and 6.20 below.

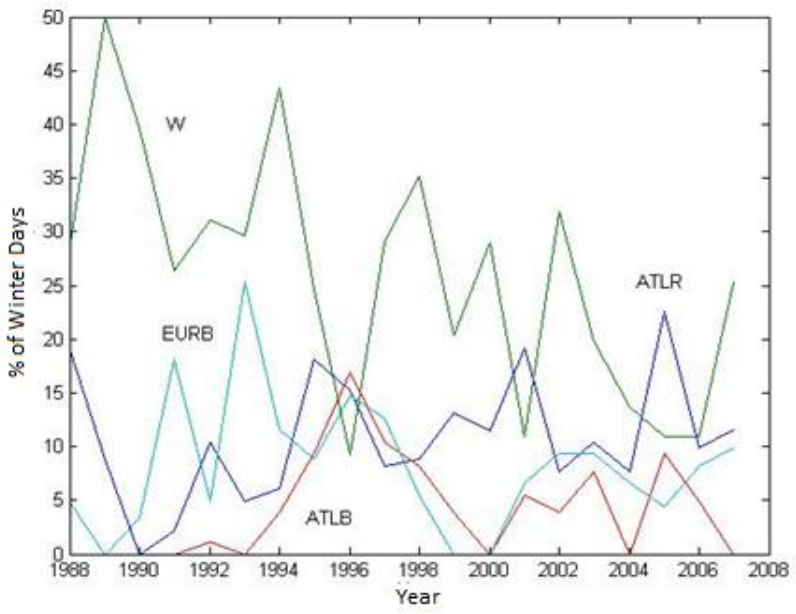

Figure 6.19. Time Series of Annual Relative Incidences of Winter GWL types, part 1.

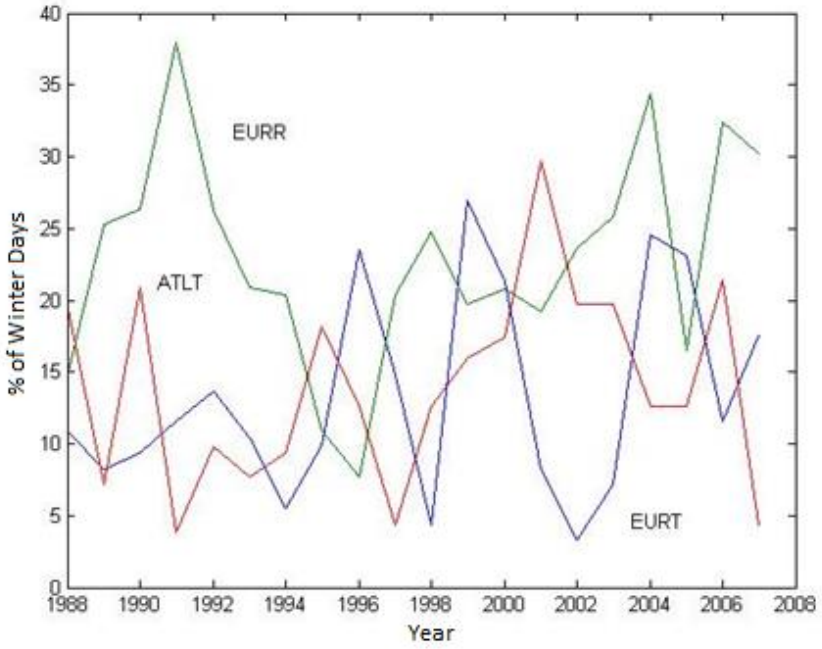

Figure 6.20. Time Series of Annual Relative Incidences of Winter GWL types, part 2.

The figures show that the circulation types are reasonably well spread out, although some are obviously more common than others. There are clearly clusters of 2 to 3 years where each type is more prevalent, but also lower-frequency oscillations with periods on the scale of decades. The same may be said for summer circulation types, therefore there is no need to present them.

6.3.2. The Effect of the Circulation Types on Principle Component Distributions

An investigation was conducted on the effect of reduced GWL circulation type on the distribution of the first 5 principal components. This was done by separating all hourly principal component values according to the reduced GWL type associated with the day in which they were recorded. The hourly values were further placed into 20 equally spaced bins, and plots produced of the relative frequency of occurrence for each bin, and each circulation type. It was found that the effect was subtle, but most pronounced for the 1st principal component, and during the winter. A plot of this is shown in figure 6.21 below, essentially a set of transparent histograms.

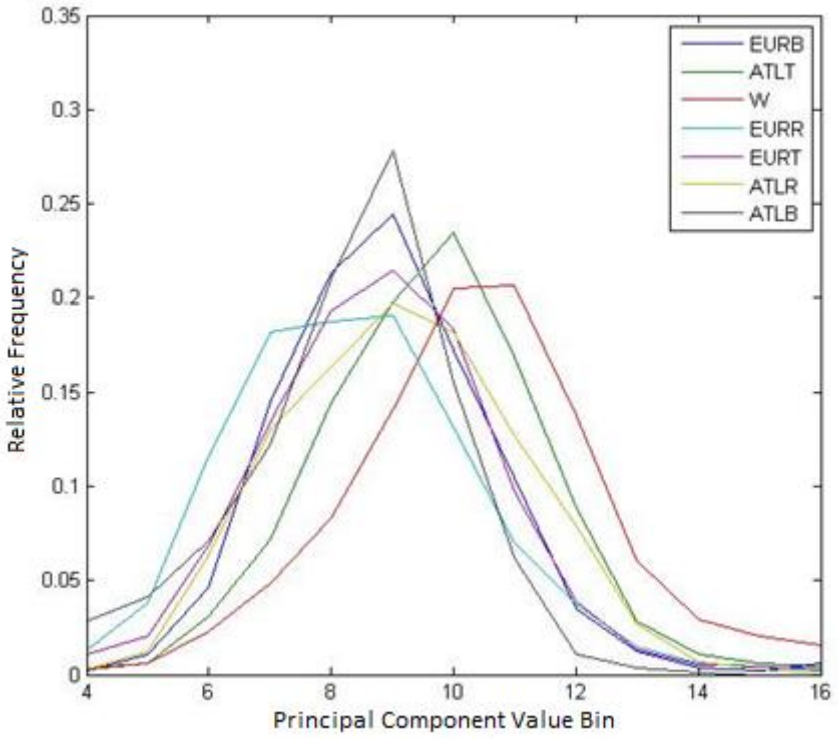

Figure 6.21. Transparent histograms for the 1st PC, for different reduced GWL types, winter.

The figure shows that the GWL is far from being uniquely determined by the 1st (or any other) principal component. However, if the principal component is within the ranges 4 -7 or 11 – 16 it is significantly more likely to be in certain GWL states than others. All GWL types have low frequencies of occurrence for values 1-3 and 17-20, which is why these bins have been omitted from the figure. So, it has been established that PCA allows the wind speed field to be related to the general meteorological 'situation' only under certain conditions.

6.4. Wind Speed Field Clustering

In addition to PCA, the value of clustering the wind speed field was investigated. The motivation was that for any given hour the wind speed field may be related to its 'nearest' cluster – i.e. the minimal sum of squared wind speed differences from the cluster centres. It was hoped that each cluster may be associated with particular types of weather – or at least much increased/ reduced probabilities of certain weather types. Previous useful wind field clustering work was reported in Chapter 2.

To make the task easier, the number of zones was reduced to 5, then gradually increased to 10, with the clusters associated with several combinations of zones explored – to find the size which gave the most spatially interesting set of clusters. In this context, 'interesting' means that the clusters display clear patterns such as a north-south or east-west divides, or naturally seem to suggest some familiar weather type. The clustering was on wind speeds aggregated into daily means, with the long-term means subtracted, such that the most transient noisy patterns were not considered. Clustering was carried out through a k-means algorithm by a single *Matlab* function. It was decided that there should be in the region of 6 clusters, reflecting the number of reduced GWL states.

A simple method for examining the results within excel was developed. This involved roughly representing the shape of GB within a rectangle of cells, with cells chosen within it that approximately correspond to the locations of the 20 Met Office stations. For a given sub-set of stations, once the set of mean wind speeds for each cluster was established, they were placed in the appropriate locations in the rectangle. The cells were then also coloured according to how windy the locations are. An example of two good, contrasting clusters is shown in figure 6.22 below.

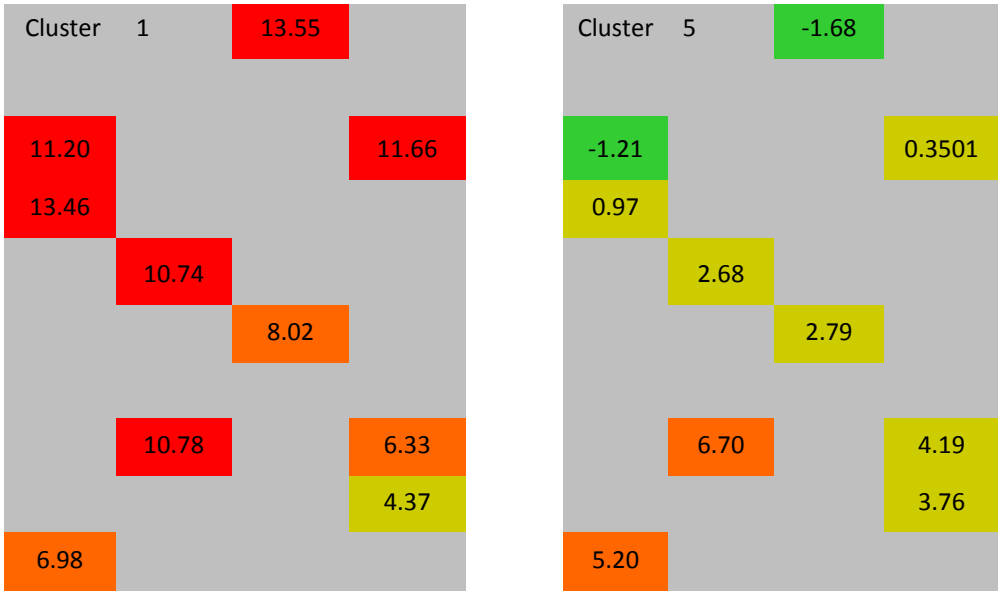

Figure 6.22. Two contrasting examples of clusters (from a set of 8), for a 10-zone representation of the (zero mean) wind speed field, in knots. That is, each value represents the wind speed at a recording station for the cluster. The relative positions of the cells roughly reproduce the spatial arrangement of the corresponding recording stations within GB

The examples shown in figure 6.22 are from a clustering scheme involving 10 locations and 8 clusters. This combination yielded results that seemed more varied than other, similar numbers of clusters. On the left is the 1st cluster, occurring on about 9% of days and much more common in winter. Winds are universally high, but Scotland is the windiest area. On the right is the 5th cluster – an overall windy state, but with the south west of England windiest, and the north and west of Scotland calm. This cluster occurs on about 17% of days, more often in summer than winter. Unfortunately, some of the other states are less spatially coherent, and discussion with Dr David Brayshaw (University of Reading meteorologist) lead to the conclusion that they probably do not represent a meteorologically meaningful way of categorising days. As a result, it seems that this is not an avenue of investigation warranting further investigation. It was suggested by Dr Brayshaw that this might not be the case if wind directions were also being simulated.

6.5. A Possible Methodology for Connecting the Wind Speed Field and Electricity Demand

The fact that the value of principle components at a given hour may provide significant information about the likely GWL state(s), as discussed in 6.3.2 above, suggests the possibility of a methodology for generating accurate coincident synthetic time series for the wind speed field and system electricity demand. Such a methodology is discussed merely in principle here, but would be a vital component in the sequential Monte Carlo simulation of the entire power systems.

The concept is that once a synthetic wind speed dataset has been generated, one could then generate a coincident daily time series of GWL types, based on e.g. a semi-Markov model. Such a model would generate states according to its own dynamics, but with the additional feature that states generated are initially tentative and are checked for compatibility with the wind series' 1^{st} PC, and probably also the next few components if the 1^{st} has a central value. A uniform random number could be generated, and if it is smaller than a certain threshold value, related to the conditional relative occurrence of the GWL types, the tentative GWL state is rejected and the algorithm must 'try again'. In this way, the GWL states would not be determined by the wind speed field, but would be consistent with it. Separate probability distributions and temporal correlation structures could be established for total demand for each GWL type, conditional of course on the time of year, day of the week and time of day.

The relationship between wind and demand would not be reproduced perfectly by such an algorithm, but it seems that the methodology would be a novel development and a valuable contribution to the field of adequacy assessment. A second use for such a scheme would be to compare the very low frequency dynamics of the series of reduced GWLs produced in this way with the long subjective GWL catalogue described in Chaprer 2 (2.3.1), as an important additional validation of the wind model.

6.6. Chapter Summary

The first results presented in this Chapter showed that long-term mean wind speeds vary across the country as expected, and that windier sites also have greater variance. There is a fairly weak correlation (coefficient of 0.23) between the mean and variance of individual months. The power transformations brought skewness values closer to zero, but many became negative due to the smaller peak to the left of the main one. Individual months are generally very different to each other, with their distribution shapes often significantly different.

Annual means show considerable variability, displaying apparently stochastic oscillations with periods of decades or more. Monthly means also appear stochastic, but have a clear periodicity of 12 months. Monthly variances follow a similar dynamic, which is partially stabilised by the power transformation.Cross-correlations were examined. For zero lag they range from 0.8 to 0.08, and none are worryingly less correlated to the others. Cross-correlations do not peak at zero lag, presumably due to a prevalent weather direction, fronts in particular. Partial-autocorrelograms are consistent with the results of previous modelling work, i.e. all models will be close to persistence, but going up to a high order does bring small additional benefit.

Long autocorrelograms and periodograms are certainly consistent with long memory, and also provide evidence in favour of modelling the series as a Gagenbauer process – although some zones more so than others. It is clear that a model which assumes constant parameters will be limited in its ability to reproduce the behaviour of wind. However since a multivariate model with such a high number of dimensions inevitably has a very large number of parameters, it would be infeasible to fit more than one set.

The set of series were transformed into their PCs. The 1st PC was found to dominate, accounting for 43.71% of total variance, and is essentially a spatial average. Occurrence rates for the reduced GWL circulation types discussed in Chapter 2 were explored and found to form clusters of years in which they occur more often, with even lower frequency dynamics also present. It was found that the circulation type has some fairly subtle influence on the distributions of the PC's values, particularly for PC1 in winter. For central and extreme values of PC1, little can be said about the corresponding reduce GWL, but for 'shoulder' values some GWL types are much more likely than others. This observation lead to a concept that could be explored in the future about how to generate coincident synthetic series for the wind speed field and demand.

Clustering of the wind speed field (K-means) was also explored as a possible way of connecting wind speed and demand. Although the results were interesting, this is not seen as a good direction for future work.

Chapter 7

Fitting a Wind Speed Model

Given the results of the data analysis presented in Chapter 6 it was decided that, despite some ambiguity, the best type of model for capturing the relevant behaviour of the wind speed field is a vector Gagenbauer, or VGARMA process – only first proposed in the literature in 2010 by Diongue [156], primarily aimed at the modelling of financial markets. This had to be combined with a suitable model for conditional variance and a choice of distribution for the unconditional errors – but rather than try to anticipate these, decisions and subsequent model fitting took place following examination of residuals derived from the VGARMA model. The final result was an APARCH structure, combined with deterministic seasonality, so that the complete model can be described as VGARMA-APARCH.

This represents not only a novel way of modelling wind speed, but also a novel type of time series model: while GARMA models have previously been combined with GARCH-type error structures, the Author is not aware of any reports within the literature of this being done for a multivariate process. This potentially naïve combining of models involved, admittedly, a slight mathematical leap of faith – but this did not seem to cause any problems in practice. Unfortunately, since this research project is concerned with a very specific practical application, time did not permit exploration of the mathematical properties of this new type of model beyond the extent to which it succeeds to reproduce the essential characteristics of the wind speed field. Additionally, the very high dimensionality of the current problem meant that full maximum likelihood estimation was unfeasible, so that no observations could be made on the properties of maximum likelihood estimators for the new VGARMA-APARCH process.

7.1. Fitting an Annually Seasonal VGARMA model

7.1.1. Initial fitting using Levinson-Durbin and OLS Methods

As stated above, the initial assumption was that the 20 wind speed time series should be modelled as a VGARMA process, with the annual seasonality corresponding to their Gagenbauer frequencies, and the errors assumed to be without any kind of structure, at this stage.

The initial step was to find first estimates for the differencing parameters δ_i for each series individually. This involved selecting trial values for the δ_i then fractionally differencing the i^{th} series consistent with them, using the Gagenbauer polynomial series expansion for the back-shift operator given by equations 3.30 – 3.32 in Chapter 3. The infinite series were truncated at the 100,000th term – a high and quite computationally expensive value, going back more than 10 years into the series. The large value is motivated by the fact that Gray et al., when introducing GARMA processes in [100], used an even higher value of 290,000. Once differenced, univariate AR models of different orders were fitted to the series using the Levinson-Durbin algorithm, and the lowest BIC value found. This process was repeated for an initially broad sweep of δ values, from 0.05 to 0.45 in 0.05 increments, and the value leading to the smallest possible BIC established. To ease the computational burden, syntax was used that allowed for parallel *Matlab* computation, with 8 virtual 'workers' available on the University of Bath's terminal servers. Guided by the first sweep, more detailed searches in increments of 0.01 were carried out, deemed an acceptable final resolution.

ACF's were plotted for the differenced series, up to a lag of 10,000 hours, as a check that the BIC-minimising values are large enough to completely remove the long memory. In some cases it was found that some long-memory was still present i.e. the ACF was regularly outside the white noise boundary for long lags, but an increase of 0.01 or 0.02 in δ was sufficient to resolve this. The resulting estimates indicate that long-memory is light, ranging from δ = 0.1 to 0.16, with 14 zones' in the narrower range 0.12 – 0.14. The range for the optimal auto-regressive order p was 3 – 8.

In the next stage, 20-zone VARMA models were fitted to the differenced series using the Hannan-Rissanen procedure described in Chapter 3. The Matlab Optimisation toolbox function *lsqcurvefit* was used for OLS fitting, with initial values obtained from the multivariate Levinson-Durbin algorithm. The problem was simplified to 20 separate optimisation problems, namely the minimisation of error variance for each zone individually. This means that the trace, rather than the determinant of the multivariate error matrix was minimised, under the

assumption that the coefficients will be very similar for both. Since the conditions for stationarity and invertibility are defined in terms of the determinants of the autoregressive and moving-average matrix polynomials, these were verified post-calculation, rather than included as optimisation constraints. Additionally, the matrices were verified as being positive-semidefinite, so that the process may be (somewhat crudely) approximated as a MVN if necessary.

To improve accuracy, the Hannan-Rissanen procedure was enhanced slightly. While innovations were initially estimated from an AR(10) model, several iterations or the entire procedure were executed, whereby the approximated innovations are replaced by improved ones from the last iteration, until the BIC value no longer decreases noticeably. On a few occasions the BIC began to increase after a few iterations, so clearly the process was repeated from the beginning, whilst being sure to stop at the established point of minimum BIC.

It was found that a mixed ARMA model is definitely the best choice, with ARMA(3,1) and ARMA(3,2) close contenders. Examination of the ACF for the residuals, up to a lag of 50, showed that for several zones there were a few correlations outside the white noise boundary for the former, but almost none for the latter, so ARMA(3,2) was chosen as the best fit model.

The final stage of the fitting was to investigate whether the univariate model δ-values needed adjustment. So, for each zone i in turn, the δ_i value was adjusted by -0.05 to +0.05 in 0.01 increments and the BIC's of the associated best fitting ARMA(3,2) models found, using the extended Hannan-Rissanen procedure. This lead to a change of ±0.01 or ±0.02 in δ_i in about half of the zones. Obviously, a completely thorough search would change the δ_i values of many combinations of zones simultaneously, but this was computationally unfeasible. As a compromise, the search was carried out twice, in case changes made to zones 2 – 20 on the first iteration had any effect on the optimum value for zone 1, for example.

Having completed estimation in this way, the next stage was to attempt estimation using maximum likelihood techniques. In order to facilitate presentation of this aspect of the model fitting, an overview of the relevant theory and previous work is given in the next section.

7.1.2. Exact and Approximate Likelihood Functions for VGARMA Processes

Diongue and Guégan in [151] report on the maximum likelihood fitting of the univariate multi-factor GIGARCH process – a GARMA process with multiple spectral poles, and with GARCH structured errors. Such a process was first proposed by Guégan in 2000.

They state that simultaneous estimation of all parameters can be achieved using conditional sum of squares (CSS). Analytical expressions are presented for the log-likelihood, which depend upon an assumed parametric distribution for the errors. By far the simplest expression is for Gaussian errors, where the log likelihood is given by

$$L(w) = \sum_{t=1}^{T} l_t/T \qquad l_t = -1/2 \log(h_t) - z_t^2/2h_t. \tag{7.1}$$

An expression is also given for the Student-t distribution, which is similar but more complex and involves the distribution's (single) parameter.

Conditions for the existence of good CSS estimators are provided, when the error structure has a GARCH(r, s) structure. In addition to the usual constraints on the AR and MA polynomials and the differencing parameters, the GARCH coefficients α_i and β_i are subject to

$$\alpha_0 > 0, \quad \alpha_1, \dots, \alpha_r, \beta_1, \dots, \beta_s \geq 0, \quad \sum_{i=1}^{r} \alpha_i + \sum_{i=1}^{s} \beta_i < 1, \quad E(z_t^4) < \infty. \tag{7.2}$$

The authors present an alternative estimation method, based on a widely used approximate quasi-likelihood function proposed by Whittle and discussed in the context of long memory process by Beran [86]. For a time series sample with a periodogram given at Fourier frequencies ω_n by $I(\omega_n)$, and associated stochastic model for the process $\{X_t\}$ with parameter vector Ψ and theoretical spectral density function $f_X(\omega; \Psi)$, the quasi-likelihood function to be minimised is

$$Q_{ML}(X, \Psi) = \sum_{j=1}^{n} (I(\omega_n)/f_X(\omega_n, \Psi)). \tag{7.3}$$

Diongue and Guégan divide the estimation of a GIGARCH process by Whittle's method into two parts: separate estimation of the GARMA and GARCH structures, and this was followed in this project. This involved estimation of the GARMA model, establishing the series of residuals and treating it as an ARMA process to be estimated using Whittle's estimator. This requires a new definition of the GARCH process structure, with z_t^2 as the response variable and a new explanatory variable $v_t = z_t^2 - h_t$, which is a zero-mean white noise process.

When the estimation process proceeds in this way, i.e. two fundamental stages, it is not essential for the same maximum-likelihood based method to be used in both stages. For example, it is possible to estimate the GARMA structure using Whittle's method, and the

GARCH structure by CSS. One method of making CSS GARCH structure estimation easier is variance targeting estimation (VTE), as explained by Francq et al. in [157]. VTE relies on a (different) re-parameterisation of the model in which the intercept α_0 is replaced by the estimated unconditional variance – i.e. sample mean, leaving the remaining two parameters to be estimated in a second step. Diongue and Guégan state that even if the sample variance converges to the population variance, the use of a two-step procedure should deteriorate the asymptotic precision of estimates, and this is shown to be the case. It is also shown however that when the model structure is miss-specified, the VTE can be superior.

Diongue and Guégan consider first the GARCH$(1,1)$ model before generalising to GARCH(p,q). The discussion reported here is for the simpler case. Recalling equations 3.34 and 3.35, a GARCH(1,1) model has the form:

$$z_t = h_t^{1/2}\,\xi_t\,, \qquad E\left(\xi_t^2\right) = 1 \text{ and } \quad h_t = \alpha_0 + \alpha_1 z_{t-1}^2 + \beta_1 h_{t-1}. \tag{7.4}$$

Thus, if the process z_t were assumed stationary, its variance would be

$$\gamma_0 = \alpha_0/(1-\alpha_1-\beta_1) = \alpha_0/\kappa_0. \tag{7.5}$$

This enables the re-parameterisation of the defining equation as

$$h_t = h_{t-1} + \kappa_0(\gamma_0 - h_{t-1}) + \alpha_1(z_{t-1}^2 - h_{t-1}), \tag{7.6}$$

which allows interpretation of κ_0 as the speed of mean reversion in variance. Writing it as the more familiar

$$h_t = \kappa_0\gamma_0 + \alpha_1 z_{t-1}^2 + \beta_1 h_{t-1}, \quad \text{with} \quad \kappa_0 + \alpha_1 + \beta_1 = 1, \tag{7.7}$$

the volatility at time t may be interpreted as a weighted average of the long-run variance, the square of the last return and the previous volatility. In this formulation, κ_0 is the weight of the long-run variance.

In addition to reducing the number of parameters, this re-parameterisation facilitates the beginning of likelihood function calculations at time $t = 1$ by assuming that

$$h_0 = z_0 = \gamma_0 \text{ (and therefore } h_1 = \gamma_0). \tag{7.8}$$

Diongue and Guégan in [158] examine how good the CSS and Whittle methods described above are for estimating GIGARCH processes, using Monte Carlo simulation. In addition to the CSS likelihood expressions for Gaussian and Student-t distributions, those for GED and skewed student-t distributed noise processes are included. They found that CSS generally outperforms the Whittle method, but that both do well, and that initially ignoring the GARCH structure does not have much influence on the estimation of the GARMA parameters. When using the Whittle approach, it is stated that the distributional parameters

(such as degree of freedom, in the Student-t case) are obtained by applying a maximum likelihood method to the standardised residuals of the GARCH model. Another paper by Diongue, Guégan and Vignal [159] describes the fitting of the GIGARCH model structure to electricity market spot prices, and the 2-step procedure is again adopted.

As stated at the beginning of this chapter, the best choice of GARCH structure and error distribution function was considered an open question during the estimation of the GARMA model. Therefore, in addition to literature on the fitting of GIGARCH processes, literature on the fitting of other conditional variance models is of interest. In [160], Peters compared the forecasting performance of several GARCH-type models for the returns of stock indices. He found that noticeable improvements can be made when using an asymmetric GARCH and that the use of non-normal densities is a promising area, as long as used with very clearly non-normal series.

Asymmetry in conditional variance known as the leverage effect. It is a common feature of financial series, and requires the APARCH structure - presented in section 3.2.2, equation 3.38, to account for it. In [161], Diongue and Guégan are concerned with modelling the price of electricity on a spot market, a series which they note displays a significant leverage effect, in addition to seasonal long memory. A new model type was therefore proposed by them in [161], which combines seasonal fractional differencing with an APARCH structure for innovations, i.e. the GARMA-APARCH model. This widens the scope of the model structure that could be chosen if modelling wind speeds at a single location.

Diongue and Guégan prove the existence of stationary solutions in [161]. They also found that the error distribution was leptokurtic for the series of interest, and so chose Student-t rather than Gaussian unconditional errors. They refer to the Ganenbauer aspect of the model as “pseudo-seasonalities associated with persistence”. This is a slightly different interpretation to that of Bouette et al. [90], described in previous chapters, with the former being more appealing. Parameter estimation is conducted through essentially the same 2-step procedure, but here with the GARMA structure estimated using the Whittle method, and the APARCH structure using CSS on the residuals series.

Diongue, Guégan and Wolff concern themselves in [162] with the exact maximum likelihood estimation of another variant of the GARCH structure, the BL-GARCH – described in Chapter 3 and set out in equation 3.36. They state that such models have been shown to be capable of providing an arbitrarily close second-order approximation to a general class of

underlying nonlinear process and can generalise various GARCH-type models, including APARCH, under certain modifications. Necessary conditions are presented for the positivity of the conditional variance, and for the stationarity of the process, and these are more demanding than for APARCH. This makes the BL-GARCH model less attractive as the constraints would have to be added to the optimisation procedure, making it more computationally challenging.

The BL-GARCH structure is presented in [162], initially making no assumptions about the structure of the conditional expectation value, but later assuming an ARMA structure. In its standard form, the BL-GARCH model assumes that the conditional distribution is normal, but non-normal alternative distributions were also explored in the paper. Indeed, analytical expressions for the conditional log likelihoods are given for Student-t and GED distributions, and they have about the same level of complexity as those for APARCH.

Having discussed so far the estimation of univariate processes, this discussion now moves on to multivariate processes. In [156], Diongue proposes the vector-GARMA, i.e. VGARMA(p, d, v, q) model – where d and v are vectors, beginning with presentation of the model structure. He does not allow it to be multi-factor, but doesn't provide any particular reason to doubt the validity of a k-factor-VGARMA as a model. Conditions are presented under which the process has a stationary and invertible solution. These are simply that determinants of the AR and MA matrix polynomials have all roots outside the unit circle, as usual, and the differencing parameters have the familiar restrictions.

The spectral density of the process is presented as (with notation change and expansion):

$$f_X(\omega) = \Delta^{d,v}(\chi)^{-1}\,\Phi(\chi)^{-1}\,\Theta(\chi)\,\Sigma\,(\Theta(\chi^{-1})^{-1})'\,(\Phi(\chi^{-1})^{-1})'\,(\Delta^{d,v}(\chi^{-1})^{-1})', \qquad (7.9)$$

$$\text{where } \chi = e^{-i\omega}.$$

For multivariate processes, the Whittle quasi-log-likelihood function becomes

$$Q_{ML}(X, \Psi) = -\tfrac{1}{2}\textstyle\sum_{j=1}^{n-1} \log\left(det\left(f_X(\omega_j)\right)\right) - \pi\, tr(\sum_{j=1}^{n-1} f_X(\omega_j)^{-1}\boldsymbol{I}(\omega_j)), \qquad (7.10)$$

where $\boldsymbol{I}(\omega)$ is the sample spectral matrix of $\underline{X}_t$, defined similarly to Chapter 3, as

$$\boldsymbol{I}(\omega) = \frac{1}{2n\pi}(\textstyle\sum_{t=1}^{n} \underline{X}_t\, e^{-it\omega}).(\sum_{t=1}^{n} \underline{X}_t' e^{it\omega}). \qquad (7.11)$$

Clearly, for a 20-dimensional process and very large sample size the calculation of $Q_{ML}(X, \Psi)$ is computationally extremely expensive – particularly given that it has to be evaluated many times in order to optimise all the parameters. However, as Diongue notes, the highest

dimension of matrices involved is 20 x 20, making the optimisation much more manageable than an exact ML method.

Beran in [86] proposes a further approximation that simplifies Whittle's likelihood calculation. Returning to univariate processes, with parameter vector Ψ organised such that

$$\Psi = (\Psi_1, \eta) \text{ where } \Psi_1 = {}^{\sigma_\varepsilon^2}/_{2\pi} \text{ and } \sigma_\varepsilon^2 = var(z_t). \tag{7.12}$$

Introducing the parameter vector $\Psi^* = (1, \eta)$, Beran shows that the vector of unknown parameters η can be estimated by approximate maximum likelihood as those that minimise the function

$$\widetilde{Q_{ML}}(X, \eta) = \sum_{j=1}^{n} (I(\omega_n)/f(\omega_n, \Psi^*)) \text{ , and that} \tag{7.13}$$

$$\hat{\sigma}_\varepsilon^2 = {}^{4\pi\tilde{Q}(\hat{\eta})}/_{n} \text{ , so that } \widehat{\Psi} = ({}^{\hat{\sigma}_\varepsilon^2}/_{2\pi}, \hat{\eta}). \tag{7.14}$$

This approximation was suggested initially for fractional Gaussian noise, under the assumption that periodogram ordinates at the Fourier frequencies are approximately independent exponential random variables with expected value equal to the spectral density (even though consistency and independence are proven for a finite number of non-zero frequencies only).

Diongue's confidence that the VGARMA model structure would be valid was derived mainly from Hosoya's theoretical work in [163] on the quasi-likelihood approach in the context of multivariate long-memory time series models. Hosoya believed that the form of the quasi-log-likelihood function suggests that the long-range dependence it can deal with is not limited fractional ARIMA, which was considered first, but is applicable to a variety of models where long memory dependence is modelled by a parametric spectral density.

Other work in this area includes Luceño [164] who suggests an *approximate* log-likelihood function for VARFIMA processes based on the process' autocovariance function. He points out that the presence of AR parameters greatly complicates the estimation, which involves hypergeometric functions that must be evaluated with a truncated infinite sum. As a consequence, a rounding error is inevitable and its impacts may not be trivial, especially when the dimensionality of the data series is relatively large. The contribution of Tsay [165] is to show that the conditional likelihood function can be evaluated exactly and efficiently if the model can be expressed with the fractional differencing operator appearing 1st on the LHS of the defining equation – i.e. if the first two operators can be swapped. This is the case if either the AR matrix polynomial is diagonal, or the differencing parameter is the same for each series.

This relates to the work of Haslett and Raftery [88] described in Chapter 4, who impose a homogenous structure on the fractional differencing and ARMA parameters when modelling wind speeds recorded at 12 different Irish meteorological stations. They do so because of the "tremendous computational burden" of using the unconstrained VARFIMA model, but Tsay states that his algorithm greatly relaxes the restrictions which need to be imposed.

Würtz, Chalabi and Luksan [166] point out that code already exists in the language *R* for the parameter estimation of univariate ARMA Models with GARCH/APARCH errors, allowing for (skew) Normal, GED and Student-t conditional distributions. The Authors state that the modular concept of the software's estimation procedure can be easily extended to other GARCH and GARCH related models. However, extension to models as complex as multivariate VGARMA-APARCH must surely be very challenging.

Not many examples were found of models such as VGARMA-APARCH being estimated in a Bayesian framework. One example is Pai and Ravishanker[167], who are concerned with the estimation of univariate ARFIMA processes. They perform Bayesian inference using Markov chain Monte Carlo methods and derive a form for the joint posterior distribution of the parameters which they claim renders estimation computationally feasible through repetitive evaluation within a modified Gibbs sampling algorithm.

Reisen et al. in [168] discuss the estimation of fractionally integrated processes with seasonal components, much less general than GARMA, and given by

$$(1-B)^d(1-B^s)^D X_t = z_t. \tag{7.15}$$

This is similar to the process described by Chapter 3's equation 3.33. In order to estimate the fractional parameters, Reisen et al. propose several estimators obtained from the regression of the log-periodogram on different bandwidths selected around and/or between the seasonal frequencies. They also consider several previously proposed semi-parametric methods and maximum-likelihood estimates and, through Monte Carlo simulations, show that the performance of their simpler estimators is good even for small sample sizes. This suggests that, in general, periodogram slopes should not be disregarded as reasonable estimators if more rigorous methods run into problems.

In [169], Tse proposes a model that extends the APARCH structure to a process that is fractionally integrated. This is known as the fractionally integrated asymmetric power autoregressive conditional heteroskedasticity (FIAPARCH) model. He limits himself to the univariate FIAPARCH$(1, d, 1)$ model, applied to the residuals of time series models with

structures that are not discussed. For the financial datasets with which Tse is concerned, the hypothesis of long memory presence is rejected by standard tests. Despite this, as the estimated values of $\alpha + \beta$ are quite close to one, Tse felt a need to examine closely the possibility of long-memory persistence in variance, thus fitting a FIAPARCH$(1, d, 1)$ model through quasi-maximum likelihood. Analysis showed however that there are no substantial differences between the stable and the fractionally integrated models. For the sake of simplicity, it will be assumed in this project that there is no need to include long memory into the modelled conditional variance structure.

7.1.3. Attempting to Estimate the VGARMA Model Parameters through Quasi-Maximum Likelihood Methods

The next stage in the model fitting process was to attempt to find superior parameter estimates based upon methods/ conclusions in the literature described above. It was clear from the outset that exact maximum likelihood would be infeasible, rather that Whittle type quasi-maximum likelihood estimators should be used, implemented through Matlab's Optimisation Toolbox.

It was also clear from the outset that even with the much simplified estimator, calculation of the log-likelihood for all 20 zones simultaneously is computationally prohibitively expensive. Upon testing the evaluation of the 20-zone log-likelihood for a *single set* of parameter values, only a small percentage of the calculation had been completed after several hours on a desktop computer with typical memory for home/office use. Given that optimisation may involve hundreds of repetitions of this calculation, it was clear that the computational expense amounted to many thousands of hours of computations. Although the nature of a PhD research project could potentially allow for such long computations, there was insufficient evidence about the suitability of the model structure to justify undertaking this course of action.

The decision was consequently made to attempt the quasi-maximum likelihood estimation for subsets of the zones – i.e. fit 4 VGARMA models for groups of 5 zones. Groups were chosen as spread evenly throughout GB, rather than regional clusters. The assumption was maintained that $p = 3$ and $q = 2$ are the best choices, even with the reduced number of zones. The motivating assumption was that the δ values established in this way are good estimates for the full 20-zone model, and that they should be used to fractionally difference

the series, facilitating estimation of the remaining parameters through the extended Hannan-Rissanen procedure. In order to establish whether δ values truly are independent of the number of zones in the model, to a good approximation, it was decided that models should first be fitted for merely pairs of zones.

In order to gain experience of optimisation with Matlab, it was decided that univariate Gagenbauer models should first be fitted using the Whittle-type estimator, and using purely AR models ($p = 4$) for additional simplicity. Mathematically the problem is one of unconstrained smooth nonlinear optimisation. Whereas constraints exist, it was assumed once again that since good starting values are available, falling easily within the constraints, the optimal estimates will be close and therefore will naturally satisfy constraint conditions. Whenever sensible results were obtained, they were checked to see if all constraints were indeed met, and this always turned out to be the case.

Matlab's Optimisation Toolbox provides two solvers for this type of problem: *fminunc* and *fminsearch*. The former may utilise one of two optimisation algorithms: 'large-scale' and 'medium scale'. The large scale optimisation algorithm is a subspace trust-region method and is based on the interior-reflective Newton method. In order to use the large-scale algorithm, one must supply an analytical expression for the gradient of the objective function, while an analytical Hessian is helpful if available. The medium-scale optimisation uses a quasi-Newton method with a cubic line search procedure. An analytical gradient is optional here, and there is no value in providing the Hessian. The nature of the objective function in our case means that it is not clear whether analytical expressions can be found for the partial derivative of the function with respect to all parameters, so only the medium-scale algorithm was considered applicable.

The *fminsearch* solver can utilise only one optimisation algorithm, which uses the simplex search method - a direct search method that does not use numerical or analytic gradients. The toolbox's user guide suggests that it may be less efficient at optimising complicated functions, although might be more robust in certain circumstances. For this reason, *fminunc* was used to perform the optimisation for univariate models. It was discovered that the solver could only find sensible solutions when the objective function as given by equation 7.10 is simplified as described by equations 7.13 and 7.14, and also when the algorithm's tolerance was increased to 10^{-3}.

It was found that the quasi-likelihood technique gave much smaller estimates for the differencing parameters δ – typically between $1/2$ and $1/3$ of the OLS based estimates. Evaluating the theoretical spectral intensity for both parameter sets for every Fourier frequency, it was found that the maximum likelihood (ML) parameters have somewhat larger values for lower frequencies, and smaller values for higher frequencies. The frequency at which the curves cross over – i.e. the OLS intensity funtion is larger than the ML, corresponds to a period of about 2 years for most zones.

At the Gagenbauer frequency, both have exactly the same spectral intensity, a value much smaller than found from the data (periodogram). This could be partially explained by rounding errors, but the fact that the theoretical spectra are so radically different in this regard suggests a genuine flaw in the model. These spectra, and the corresponding periodograms, are shown in figures 7.1 – 7.2 below, with the axes scaled similarly to 'regular' values in 7.1 and scaled to show the full peak at the annual frequency in 7.2.

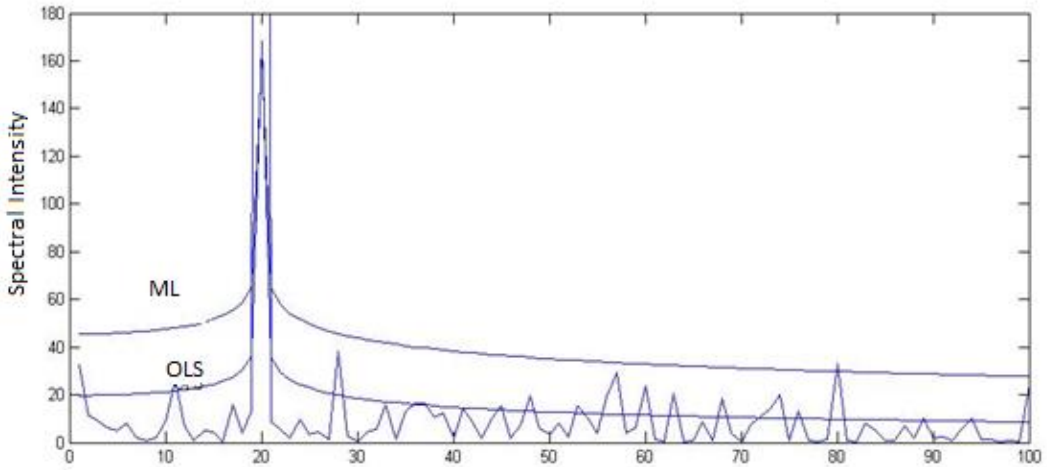

Figure 7.1. The theoretical spectral intensities for ML and OLS fitted models and the periodogram – Zone 1, lowest 100 frequencies.

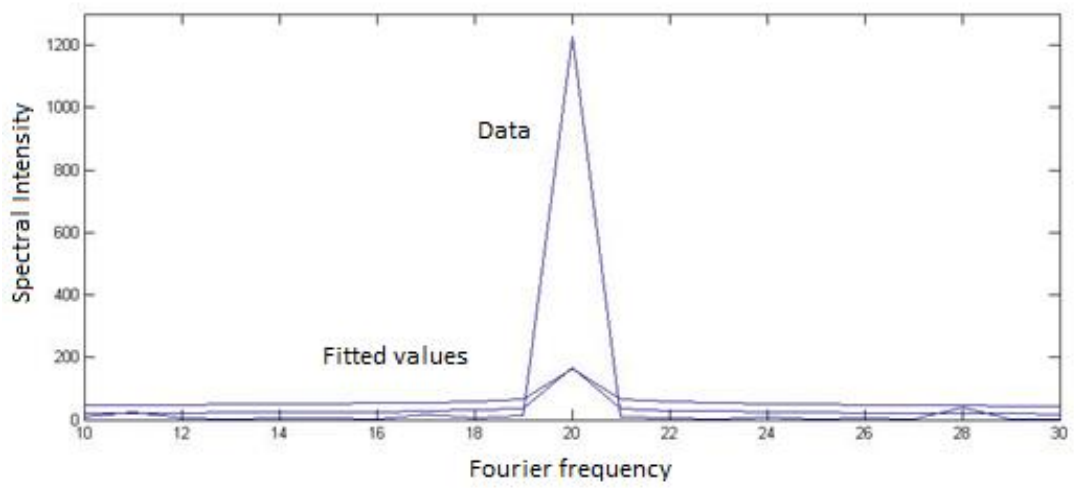

Figure 7.2. The theoretical spectral intensities for ML and OLS fitted models and the periodogram – Zone 1, frequencies surrounding the annual peak.

The OLS spectrum is clearly the closest match for the lowest frequencies. For a few example zones, a smooth best fit curve was applied to the periodogram for the range j = 40,000:41,000, which are intermediate frequencies, for comparison with the two theoretical curves. The data fit was found to lie in the middle between the ML and OLS curves for each zone. However when this was repeated for the range 80,000 to 83,000, i.e. high frequencies, the ML curve was generally a much better fit – indeed very close to perfect. This seems to suggest that for univariate models, the ML estimates are generally better, with the exception of the lowest frequencies.

Figure 7.3 shows the theoretical spectra due to the short memory part of the model only (i.e. the AR coefficients), for zone 1. The Fourier frequencies roughly correspond to $\omega = 0$ to $\omega = \pi/2$. The ML estimated spectrum is closer to that of the persistence model as a consequence of the lighter differencing the model applies to the data.

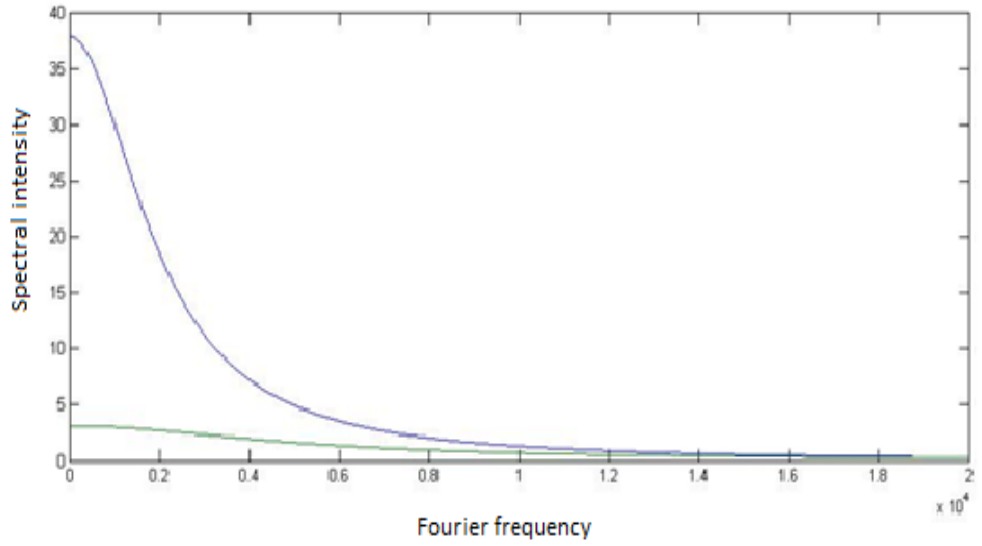

Figure 7.3. The theoretical spectral intensities for the short memory parts of the ML and OLS fitted models, Zone 1.

Before addressing concern over the univariate models' inability to reproduce the peak at the annual frequency, the optimiser was used to fit a bivariate model. Zones 5 and 9 were chosen and, using the OLS method, it was established that AR(4) was the best model order. Trying first the *Fminsearch* solver, with the tolerance again increased to 10^{-3}, the optimisation kept on going for a period of about 3 hours, when the search was terminated, without displaying any signs of converging. The objective function was in fact several orders of magnitude greater than its starting value at this point. The number of function evaluations per iteration was surprisingly small, averaging about 1.5. Moving on to *Fminunc*, the algorithm initially seemed to be finding a local minimum. With many more function evaluations per iteration, each one took a few minutes, until the 6^{th} iteration was reached – at which point the algorithm could not proceed, despite an ever-increasing number of function evaluations over a period of several hours. The optimisation was attempted again, this time with the tolerance increased to 10^{-2}, and it completed successfully on the 5^{th} iteration, after 27 minutes. Reassuringly, the two δ values were very similar to those found for the zones' univariate models.

The model with bivariate ML parameter values was applied to the data and the residuals series examined. Examination of ACFs showed that the smaller δ values are in fact insufficient to eliminate long memory from the series. Also, the residuals' covariance matrix elements were significantly larger than their theoretical values, according to the ML

optimisation – i.e. the ML parameters are not entirely consistent. The ML model residuals' covariance matrix elements were in fact larger than those for the OLS model residuals, indicating clearly that the ML estimates are not superior in this case. Without attempting to fit further ML estimates to higher-dimensional models, attention turned to the problematic peak at the annual frequency, as described in the next section.

7.2. Fitting a 2-Factor VGARMA Model

7.2.1. Removing the Wind Climate

It was noted above that the spikes in the periodograms at the annual frequency, for each zone, are much bigger than their theoretical spectral intensity counterparts. This might be partially the result of rounding errors in the latter – in the denominator of the equation, but might also be explained by the annual seasonality consisting in fact of both deterministic and stochastic components. For a finite sample with only 20 repetitions of the annual cycle, an average cycle can obviously be extracted, and it is likely to represent a significant contribution to the total variability. It may be the case that treating this mean pattern as an estimate of a stationary, deterministic seasonality - i.e. a wind climate, could lead to better modelling results. Figure 7.4 below shows daily averages for the transformed wind speed series – for 4 example years, zone 3, during roughly the 4th quarter of the year. The figure shows that the annual seasonality seems to be very stochastic in nature, and that initially treating it as such was certainly justifiable. However figure 7.5 shows the 'means of the daily means' pattern, somewhat smoothed, for the same zone. Given that the variance of the series are close to 1, the trend is significant and it is definitely worth removing all these trends and exploring the resulting series.

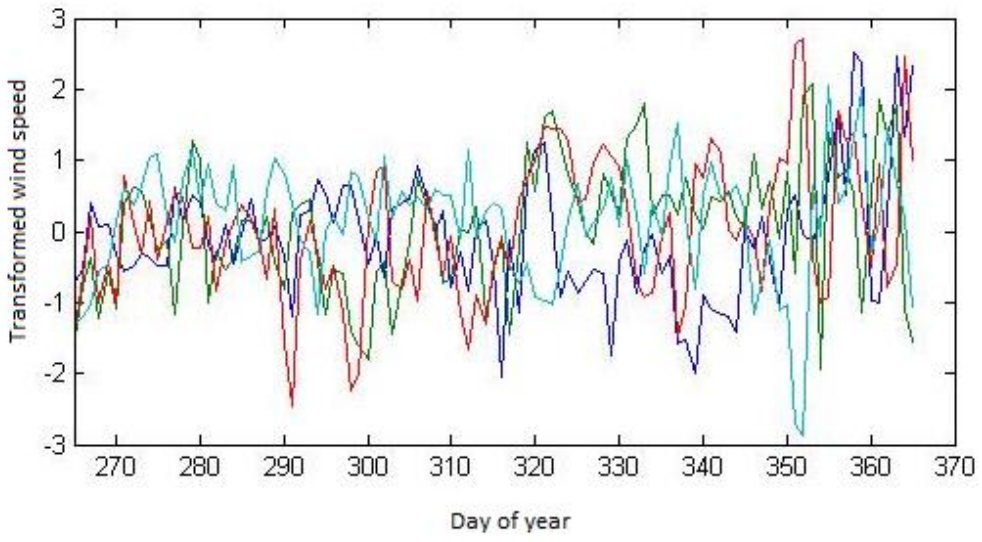

Figure 7.4. Time series of daily means, approximately last quarter of the year for 4 years, Zone 3.

It was found that the series initially generated from taking means of each year's daily mean were quite noisy, as one might expect. There was also a strong presence of several smooth harmonics – some of surprisingly high frequency and amplitude. Some smoothing was clearly necessary to eliminate the noise, but this should not eliminate too much of 'genuine' high frequency harmonics. Light smoothing seemed to work well, as can be seen in figure 7.5, which is the result of with a linear moving average filter with weights [0.2, 0.4, 0.6, 0.8, 1, 0.8…..]. Whether some of the higher frequency harmonics were genuine seemed questionable, so a heavier linear filter was applied, with weights [0.1, 0.2, …, 0.9, 1, 0.9, …]. This had the effect of dampening some harmonics, but their presence remained strong in such a way as to support the view that they are all genuine, and that smoothing should be light.

The daily average wind climate patterns, smoothed with the first filter, were stretched into hourly series through linear interpolation and removed from the hourly transformed wind series. Figure 7.5 shows that removing the climate was successful, the detrended series having only modest traces of the high frequency harmonics which were filtered out of the trend series. It was decided that these imperfections are small enough in relation to the standard deviations that the process can be considered zero mean without further adjustment.

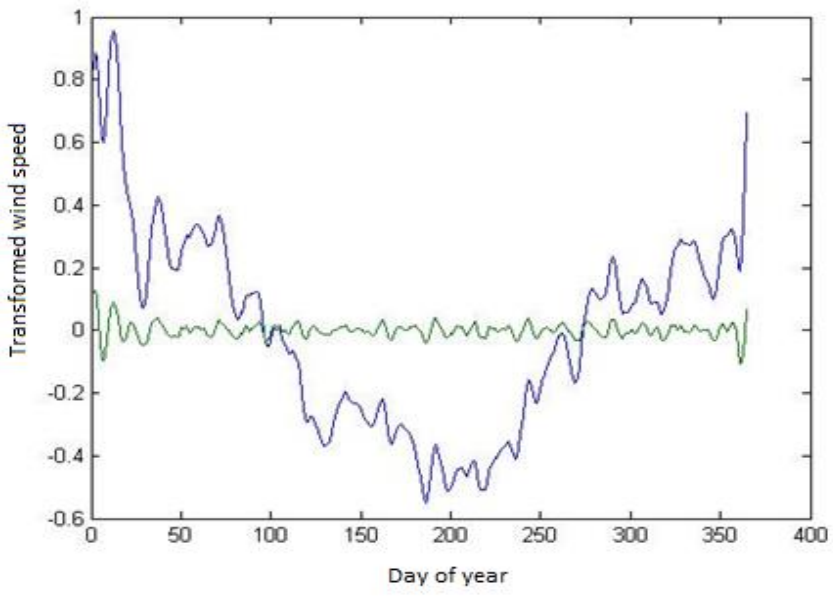

Figure 7.5. Deterministic annual seasonality for the sample period, and series with climate removed, Zone 3.

7.2.2. Analysis of the New Periodograms

Examination of periodograms for the newly deseasonalised series revealed, reassuringly, that the only change that can be observed is the disappearance of the single, very tall spike. The observed spectra may now be characterised as consisting of up to 100 spikes at the low frequency edge. Several of these spikes are now relatively large, typically in the range of about 1/3 to 1/2 of the height of the OLS fitted model's theoretical spectral intensity at the Gagenbauer frequency. An example, zone 2, is shown in figure 7.6 below. It can be seen that in this case, the tallest spike is for the lowest Fourier frequency, and it seems this zone might reasonably be modelled as a regular long memory process, with the long memory effect light. This is not the case for all zones, however – for some zones, the highest peak occurs at around the 200th Fourier frequency, for example.

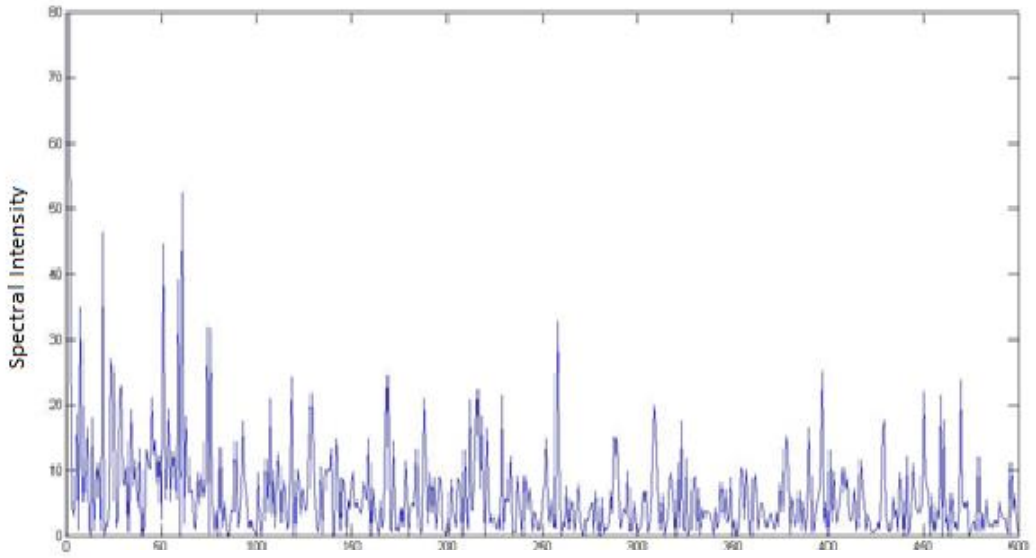

Figure 7.6. First 500 frequencies of the periodogram, entirely deseasonalised series, Zone 2.

It was decided that rather than looking at individual spikes, the peroidograms would be more revealing if heavily smoothed, so that trends and clusters of high intensity could be examined. The result of this for the same example of zone 2 is shown in figure 7.7 below. It now appears that the series might be more suitably modelled as having a peak at about the 60th Fourier frequency. However, this is far from being the only cluster present, and probably the second most prominent smaller cluster is centred at about the 220th frequency. In fact, a basic pattern was found that is almost universal across the zones: a clump of high intensity, clearly containing the greatest peak and centred between 1 and about 100 Fourier

frequencies; along with several smaller clusters. One of these smaller clusters is somewhat more prominent, and is centred between roughly 60 to 400 Fourier frequencies. It therefore seemed more accurate to assume that there are two Gagenbauer frequencies, and model the transformed wind speeds as a multivariate 2-factor-GARMA process. This is also advantageous over a single-factor model since the individual δ values will be smaller and the spectral intensity less 'peaky', i.e. more spread-out, in line with the periodograms.

For every zone, the best choice of two Gagenbauer frequencies were established by visual examination of the periodograms smoothed to various extents. The results are presented in Appendix A1.

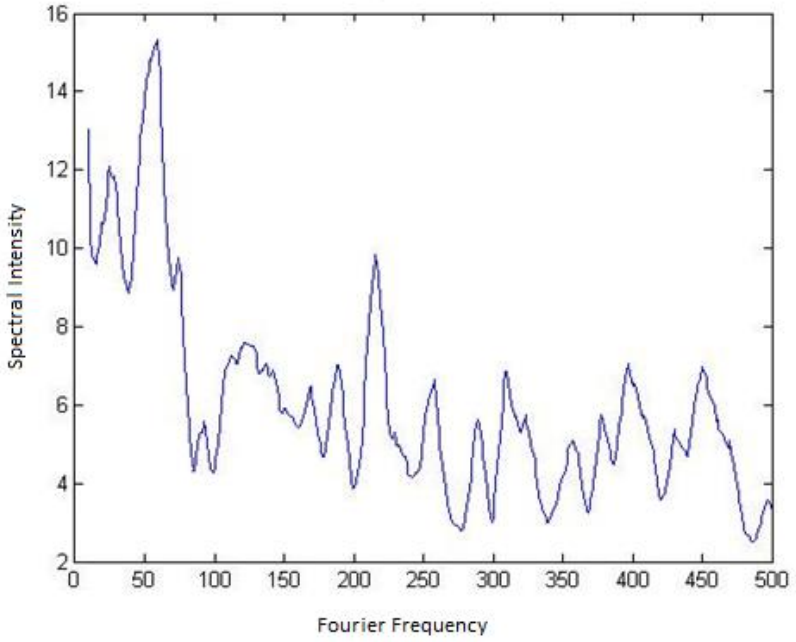

Figure 7.7. First 500 frequencies of the heavily smoothed periodogram for the entirely deseasonalised series, Zone 2.

7.2.3. Fitting the Model Parameters

The first step in fitting the 2-factor-VGARMA model was to obtain initial estimates for the two differencing parameters for each zone. This was done first for univariate models using essentially the same grid search procedure as used for the single-factor model, described in section 7.1.1. In this case however the 'grid' had 3 dimensions: two differencing parameters and the autoregressive order p, which added considerable computational expense to the problem. It was obviously not possible to truncate both differencing polynomial expansions at the previous value of 100,000 (as the entire series is of length 175,291). It was decided instead

to truncate the lowest frequencies at 70,000 and the highest at 50,000 leaving the fully differenced series at 55,291 hours, i.e. roughly 6 years and 3.5 months.

After obtaining estimates for 8 zones, it became obvious that in each case BIC values were reasonably close to minimal when the 2nd differencing parameter was 0.03. It was therefore decided that the model will adopt this parameter value for each zone. This allows the fitting to be simplified considerably whilst allowing sufficient freedom for very close to optimal solutions to be found. Model fitting proceeded with 2-dimensional grid searches that established initial estimates for the differencing parameters associated with the 1st Gagenbauer frequencies for each zone. The theoretical spectra for these univariate models were found to be significantly better fits to the periodorgams than the single-frequency models were. Plots of these spectra are not included here, for brevity, since simulation and historical periodograms are compared in the next chapter.

The next step was to investigate whether the Whittle-type ML estimator described in section 7.1.3 might be more successful for the new model structure. Both univariate and bivariate models were fitted, with AR(4) short memory structure, once again using the *Fminunc* solver and with tolerance reduced to 10^{-2}. It was found that, as before, values given for the noise covariance are not the same as those obtained from model residuals – i.e. the models are inconsistent.

Before abandoning the possibility of obtaining approximate maximum likelihood estimates, it was decided that the software *Maple* should be used to see whether a simplified analytical expression could be obtained for the likelihood, to simplify its evaluation by the *Matlab* solver. Using command-line *Maple*, calculation of the analytic expression in the bivariate case for only one discrete frequency (arbitrarily $j = 10{,}000$) was interrupted after ½ hour and an allocated memory of 32GB. The simplified expression for the theoretical frequency alone, in 1-d mathematics of font size 11, takes up 23.5 A4 pages.

It was therefore obvious that the best model fitting option is the extended Hannan-Rissanen procedure used to fit the single-factor VGARMA model. For the new model specification it was found that the best order choice for the short memory aspect is ARMA(3,1) (as opposed to ARMA(3,2) previously). The final choices of differencing parameters for the lowest of the Gagenbauer frequencies are given in appendix A1, whilst the final ARMA coefficient matrices are given in appendix A2.

It is interesting to note that the differencing parameters show considerable range, varying from 0.03 to 0.13. For those zones with the heaviest differencing, the corresponding diagonal element of the 1st autoregressive coefficient matrix tend to take values of about 0.6, while they are close to 1 for those with the lightest differencing. This relationship is not surprising, since a large value for either is a way of ensuring high spectral intensity at the low frequency extreme. This consideration casts some doubt whether the extent of long memory presence in the 20 series truly varies as much as the differencing parameters suggest, as it may rather be a quirk of the fitting procedure, similar to an anti-colinearity. Only careful comparison of the results of simulation with the historical series can shed light on this question.

7.3. Analysis and Modelling of Residuals/ Noise

The model fitting work described so far was concerned purely with modelling the transformed wind speeds' conditional expectations. The assumption made was that after suitable fractional differencing, the series may be described adequately by a linear model for conditional expectation, plus a stationary multivariate random noise process. The noise process was assumed to be joint-normally distributed with a covariance matrix identical to be that of the model residuals series. However as established in chapters 2, 4 and 6 wind is a highly nonlinear process and nonlinear effects remain present even after power transformation, removal of climate and fractional differencing. Some of these effects must be captured by a suitable model for noise heteroskedasticity, and efforts to establish the best model based upon analysis of residuals are described in this section. This means that the overall model was fitted in two parts – not the most accurate approach, but the only feasible one for such a high dimensional model. This was established as an acceptable approach in the literature review in section 7.1.2 above. Additionally, it is shown that even after accounting for heteroskedasticity the derived marginal distributions are very heavy-tailed.

7.3.1. Analysis of the Residuals Series

The 20 series of residuals obtained from applying the 2-factor-VGARMA model to the 20 year sample is the basis of noise model fitting. Before embarking on that task, it was important to examine the success of the former model as reflected in the extent to which the residuals are devoid of temporal structure. To achieve this, a 'correlation cube' was calculated, consisting of the set of correlation matrices with different temporal lags between the series pairs – ranging from zero to 10 hours. Figure 7.8 below shows a 'slice' through the cube where the temporally leading zone is kept constant - zone 9. In other words, the figure shows all auto- and cross-correlations for zone 9, lags 0 to 10 hours. This location is in Cumbria, probably the most central of locations.

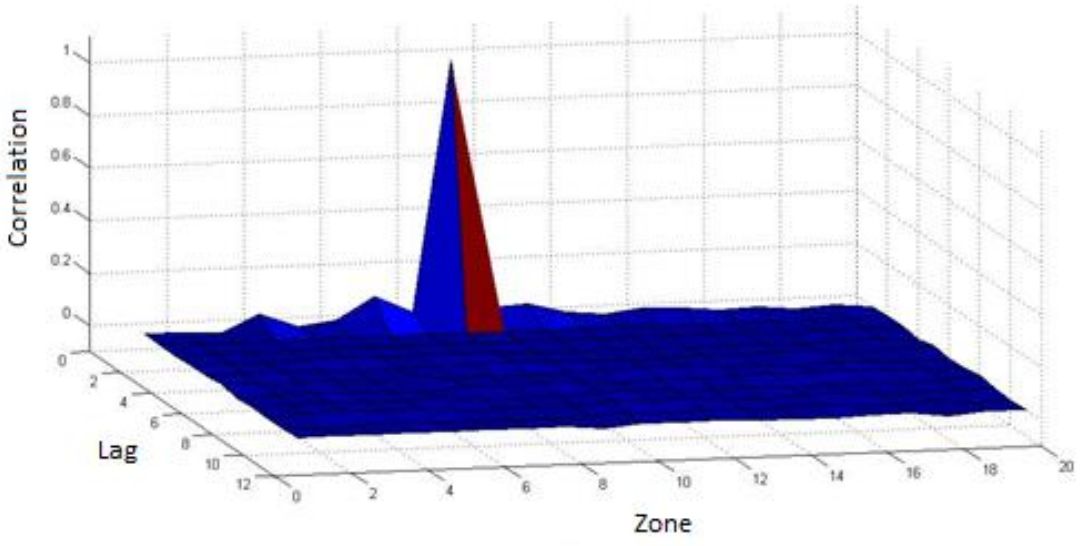

Figure 7.8. All correlations of residuals for zone 9, temporal lags of 0 to 10 hours.

As expected, the lag-zero autocorrelation is 1, and lag-zero cross-correlations (i.e. purely spatial relationships) have small positive correlations of up to about 0.1. For lags greater than zero there appear to be few, if any, significant auto- or cross-correlations, with only a few very small negative cross-correlations noticeable. This is a very positive indication that the model is a good fit.

Figure 7.9 gives a more complete picture of the spatial structure, i.e. all auto- and cross-correlations with no temporal lag. Again we see that cross-correlations are small, taking values of up to about 0.1, while auto-correlations are unity.

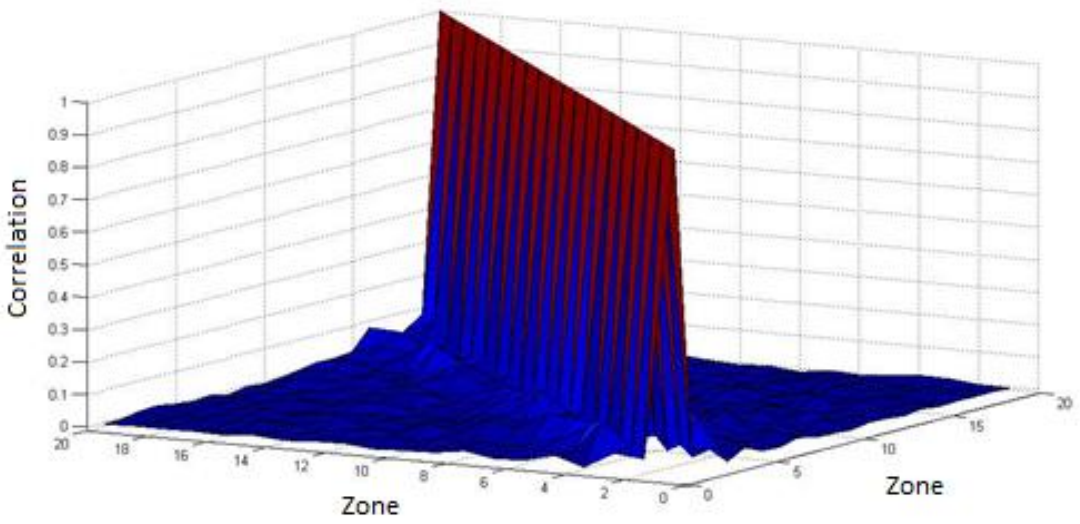

Figure 7.9. All auto- and cross-correlations of residuals, zero temporal lag.

Examining the equivalent matrix with a lag of 1 hour, it was found that all correlations are insignificant, almost all being smaller than 10^{-3}. The Gaussian white noise boundary at 95% confidence is about $2/(55287)^{0.5} = 8.5 \times 10^{-3}$, which is several times greater than any of these correlations. The same is true for a lag of 2 hours, although correlations are generally larger. Somewhat surprisingly, correlations increase considerably with increasing lag, peaking in the range 5-8 hours, and the most significant correlation, -0.0294, is found for a lag of 6 hours. This is still very small but significantly greater than the white noise boundary. For this lag, all correlations except that one lie within the range ±0.015, but 60 correlations (15%) had magnitudes outside the confidence boundary. This seems to suggest that a higher order model should perhaps have been chosen. This is however contrary to the information criteria and is not considered compelling enough evidence to justify a change. It might represent the limits of what a regression model can achieve.

Models of multivariate heteroskedasticity were discussed in Chapter 3, with equation 3.40 in section 3.2.2 stating that the individual locations' errors at time t are derived from a set of independent unit variance errors, pre-multiplied by the Cholesky decomposition of the dynamic covariance matrix. In general, the dynamic nature of the covariance matrix is such that the correlation matrix is also dynamic. If however the correlation matrix is constant, to a reasonable approximation, the situation is considerably simpler and univariate models are sufficient to describe the dynamics of variances. It is clear from Chapter 3 that since we are concerned with a very high-dimensional problem, multivariate models of heteroskedasticity inevitably involve a very large number of parameters. As a result, choosing univariate models is extremely advantageous, if viable.

In order to see whether this is the case for our residuals series, another 'correlation cube' was constructed for the series squared. Figure 7.10 is an example plot taken from this cube, showing all auto- and cross-correlations for zone 3, lags 0 to 10. It is clear that auto-correlations are certainly significant, but while several cross-correlations are significant, they are much smaller. This is not surprising, given the weak spatial structure already established. As a result, we may conclude that univariate models are probably good enough.

Plots in Chapter 6 showed that heteroskedasticity exists on timescales from intra-daily to inter-annual, with a strong stochastic annual seasonality coupled fairly weakly to the seasonality in mean. Before fitting models to capture the short-term dynamics of heteroskedasticity, it was therefore necessary to explore the seasonality of the residuals series, so that it may be removed.

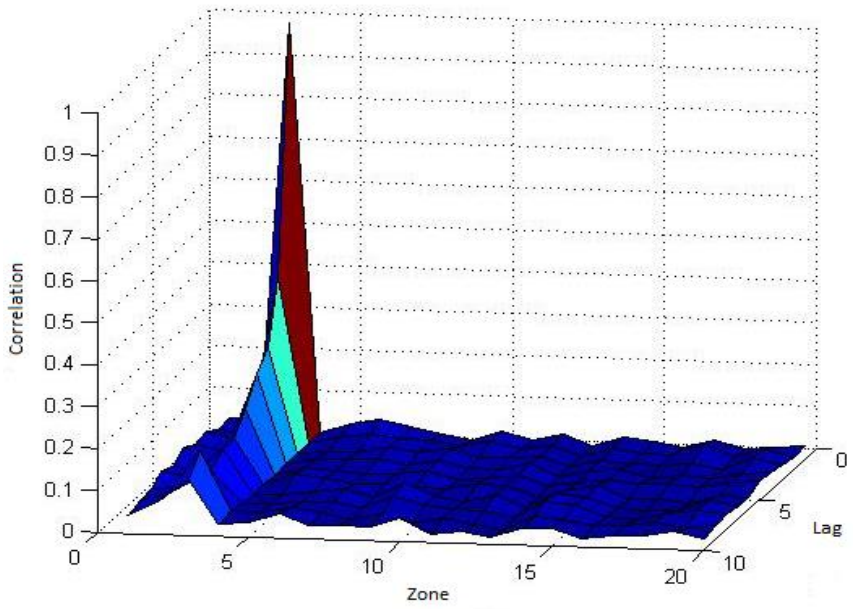

Figure 7.10. All auto- and cross-correlations of squared residuals for zone 3, temporal lags 0 – 10 hours.

To do this, daily averages of each series' absolute values were calculated, which were further averaged across the 6 or 7 years present in the series to create an annual profile. Optimal smoothing was explored and applied, following the same principles as for the transformed wind speed series. Additionally, standard deviations were calculated for the inter-annual variability of daily mean noise, for each day of the year. These were smoothed to create annual profiles of inter-annual variability – with heavier smoothing applied due to the extremely small sample sizes involved. The results are shown for the same example of zone 3 in figure 7.11.

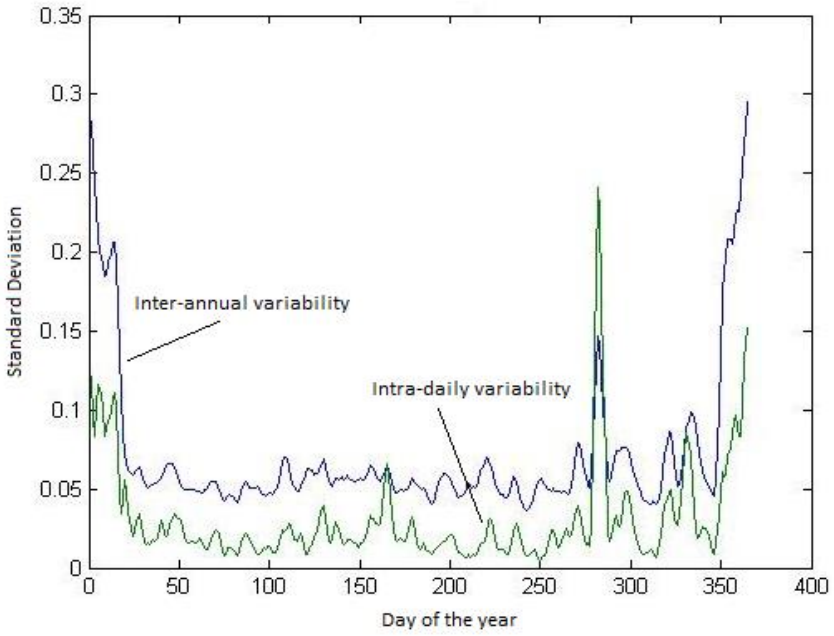

Figure 7.11. Seasonalities within the daily-averaged standard deviation of residuals, and within their inter-annual variability.

The plot shows that both profiles are naturally quite noisy, but appear to be fairly flat for most of the year, with the exception of two periods where they have significantly greater values: the 2nd half of December to the 1st half of January and roughly the first 10 days of November. Plots for most zones are similar, which is a reassurance that the patterns are real. However a similar feature to the smaller November peak is only found in a minority of zones, and a few have their significant peak during the spring.

To conclude this subsection, a simple time series plot is shown in figure 7.12 below for a typical error series segment, randomly selected zone 10, along with the series squared. A volatility clustering is very obvious, and there is a hint of a possible asymmetry, namely that large negative values might have a greater effect on the subsequent variance than positive ones.

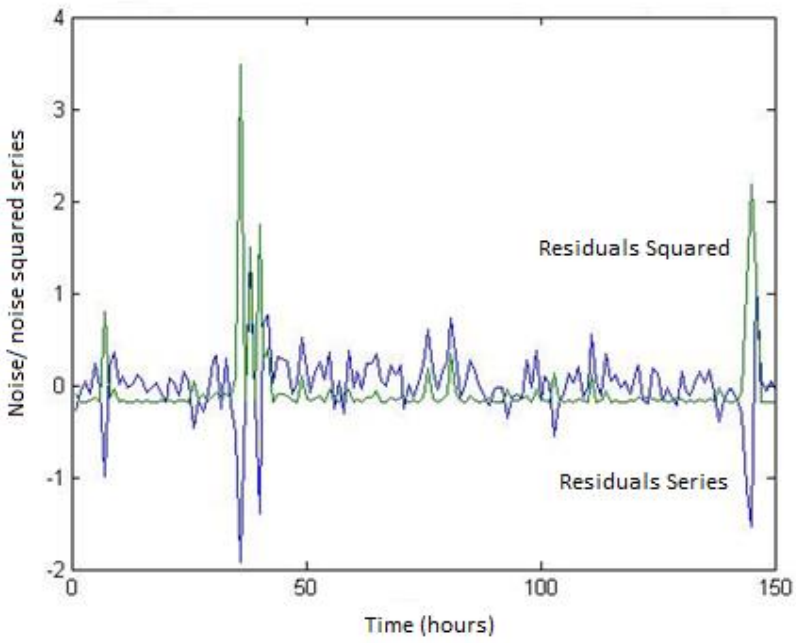

Figure 7.12. Example time series plot of the residuals and residuals squared series, zone 10.

7.3.2 Removal of Noise Seasonality

This section is concerned with the removal of deterministic seasonalities found to be present in the residuals series – both in the mean absolute value of the residuals and in the inter-annual variability of their absolute values. This was inherently a more challenging task than the removal of seasonality from the transformed wind speed series since it had to be achieved through division and power transformation only, rather than subtraction and division.

Naively, it would seem that one should simply use the seasonal mean profile for division, and the inverse of the inter-annual variability profile for power transformation. However it was found that this procedure tends to over-compensate i.e. creates new seasonal profiles that are roughly mirror images of the originals. For some zones, the mirror image profiles were actually worse than the originals, in terms of the height of seasonal spikes. As a result, several more sophisticated algorithms were explored. The structure of the most successful algorithm is given below, but before presentation a few definitions are necessary.

Let the residual at time t at the i^{th} zone be $z(i,t)$, the (daily resolution) profile of means be $m_d(i,t)$ and the (daily resolution) profile of inter-annual variability be $s_d(i,t)$, with separate profiles for regular and leap years. The procedure was:

- Calculate the ratio matrix $r_d(i,t) = s_d(i,t)/m_d(i,t)$ and apply moderately heavy smoothing through a linear moving average filter (keeping same symbol);
- Use linear interpolation to 'stretch' this matrix and the profile of means to form $r_h(i,t)$ and $m_h(i,t)$, respectively, with hourly resolution;
- Normalise these as $r_{norm}(i,t) = r_h(i,t)/\overline{r_h}(i)$ and $m_{norm}(i,t) = m_h(i,t)/\overline{m_h}(i)$;
- Power transform the rows of $r_{norm}(i,t) \rightarrow r_{trans}(i,t) = (r_{norm}(i,t))^{a(i)}$ with parameters $a(i)$, then transform the $z(i,t)$ into $z_{flat}(i,t) = (z(i,t)/m_{norm}(i,t))^{-r_{trans}(i,t)}$;
- The $a(i)$ are established through crude optimisation, based on visual examination of the end result. For 4 zones $a(i) = 0$, i.e. the variability information is lost entirely, while for the other zones it ranges from 0.1 to 0.4;

The resulting series had seasonal profiles that were quite noisy, but essentially flat, demonstrating that the method was successful.

It is assumed that removal of deterministic seasonalities in both the conditional mean and noise aspects of the wind speed process adequately accounts for the fairly weak positive correlation between mean and variance observed in Chapter 5. In other words, it is assumed without investigation that there is no need for an ARCH-in-mean element of the model structure, similar to that of equation in Chapter 3.39.

7.3.3. Fitting the APARCH Model Parameters

The next and final part of the model building process was to fit univariate GARCH-type models to the de-seasonalised residuals. In Chapter 3, a very general type of model was presented, the APARCH, and due to its generality this was chosen as the starting point here. Repeating equation 3.38, this model structure is

$$(h_t^{1/2})^\delta = \alpha_0 + \sum_{i=1}^{p} \alpha_i (|z_{t-i}| - \lambda z_{t-i})^\delta + \sum_{j=1}^{q} \beta_j h_{t-j}^\delta (h_{t-j}^{1/2})^\delta. \qquad (7.16)$$

The correlation cube described in section 7.3.1 was re-calculated for the de-seasonalised residuals and several plots similar to figure 7.10 were examined and found to be not greatly different from the previous ones. In other words de-seasonalising has had minimal impact on the volatility clustering behavior. It was concluded that the autocorrelation functions decayed rapidly enough to justify adopting the common practice of setting p and q in equation 7.16 as 1. Given the already very large number of fitted parameters it was decided to simplify matters further by setting the power parameter $\delta = 2$, so the model fitting task was reduced to that of finding $\alpha_0, \alpha_1, \lambda$ and β_1 for each zone. If it were not for the asymmetry

aspect, the model would simply be GARCH, however examination of time series segments has gave cause to believe that significant asymmetrical effects might be present and must be represented.

Despite previous failures, it was hoped that model parameter estimates could be established through maximum likelihood optimisation. As discussed in the literature review of section 7.1.2 above, the likelihood function for APARCH processes depends on the assumed probability distribution of the residuals. Whereas the simplest option by far was to assume that they are normally distributed, it was obvious from simple histograms that the residuals are roughly symmetric but very heavy tailed. Two alternative distributions were available, since the relevant likelihood functions are given in the literature review: student-t and generalised error distributions (GED). The first task was therefore to decide which of the two is the best choice overall, achieved through plots comparing the empirical probability distributions with the best fitting student-t and GED distributions, for each zone.

The empirical distributions were estimated through kernel density estimation. It was decided that the heavy but nonetheless very sparse tails would best be represented by normal kernels and, guided by [170], the bandwidth h_B was calculated according to

$$h_B = (4/3)^{(1/5)}\sigma n^{-(1/5)}, \tag{7.17}$$

where σ is the sample standard deviation and n the sample size. For centred student-t distributed random variable X, the PDF is

$$f_X(x;l) = \left(\Gamma((l+1)/2)/\sqrt{v\pi}\Gamma(l/2)\right)(1+x^2/l)^{-(l+1)/2}, \tag{7.18}$$

while if X is GED distributed, the PDF is

$$f_X(x;v) = \left(v2^{-(1+1/v)}/\lambda_v\Gamma(1/v)\right)\exp\left(\left(-\frac{1}{2}|x/\lambda_v|\right)^v\right), \tag{7.19}$$

$$\text{where} \quad \lambda_v = \sqrt{(\Gamma(1/v)2^{-2/v}/\Gamma(3/v))}. \tag{7.20}$$

The best fitting parametric distributions for both families were found by conducting grid searches of increasing accuracy for the distribution parameters, with the goodness of fit assessed visually only. It was found that student-t is probably the best choice for about 5 zones only. The alternative was found to be a reasonably good fit for all zones, albeit with some considerably better than others, and is therefore the best choice overall. One of the better fits, zone 8, is shown in figure 7.13 below.

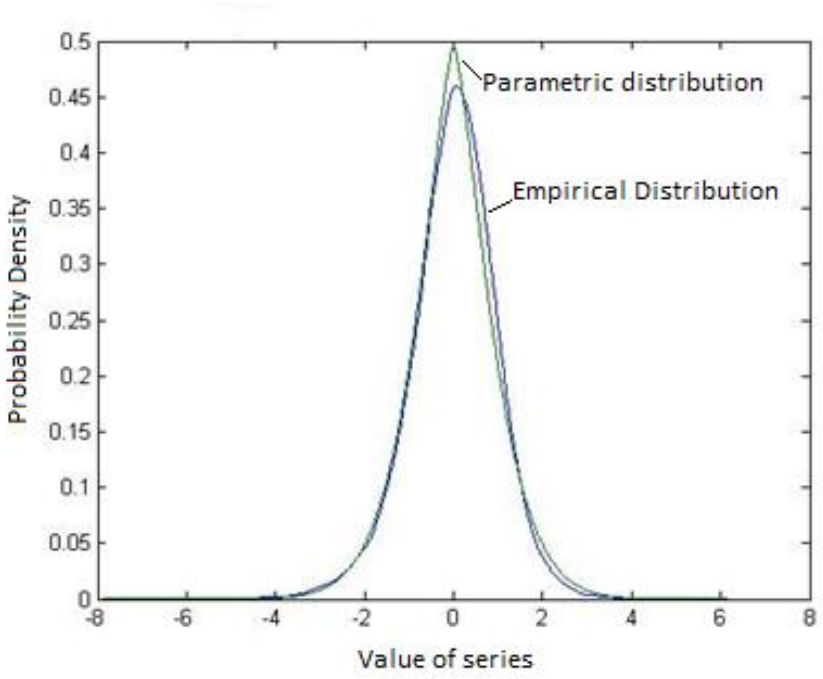

Figure 7.13. Kernel density estimate of the empirical distribution of residuals and best-fitting g.e.d. distribution.

In order to further reduce the number of parameters in the ML estimation, the VTE method described in section 7.1.2 above was adopted, leaving 20 sets of parameters $\alpha_0, \beta_0, \lambda$ and v to be estimated. Here, v is for the implied distribution of unconditional errors ξ_t, not z_t. Unfortunately the best fitting value for the original errors is the only initial estimate available for the unconditional errors, despite a probable difference in leptokurtosis. Initial estimates were also required for the other parameters before ML estimation could be attempted. This was achieved through quite computationally expensive 3-dimensional grid searches, with each parameter ranging from 0 to 1 in increments of 0.1 but with boundary conditions limiting the possible combinations. For each set of parameter values, h_t series were calculated, and from this the ξ_t series.

The autocorrelation functions for the ξ_t^2 series were then calculated, for lags 1 – 50 hours, and the sum of their absolute values used as the metric to reflect the suitability of the parameter value set. For most zones, 2 or 3 local minima were found in the 'parameter cube', all of which were assumed to be potentially valid starting points for ML optimisation.

Letting the parameters $\alpha_0, \beta_0, \lambda$ and v form the list ϑ, the associated likelihood function for g.e.d. noise, as presented in [161], is

$$L(\vartheta) = n(\log(v/\lambda_v) - (1 + 1/v) + \log(2) - \log(\Gamma(1/v))) - \frac{1}{2}\sum_{t=1}^{n}\left(\log(h_t^2) + \frac{1}{h_t}\left|\frac{z_t}{\lambda_v}\right|^v\right). \quad (7.21)$$

In the present context, it was not possible to rely on initial values to ensure that boundary conditions are met and a bound optimisation solver had to be selected in Matlab's optimisation toolbox. The appropriate solver offers a choice of four algorithms: active-set, sequential quadratic programming, trust-region reflective and interior-point. The first two do not require user-defined Hessians and since the toolbox user manual suggests that the second is generally the best of the four, it was chosen here.

For each zone, it was unfortunately found that the optimisation went on for a long time and stopped only when one of the parameter values had (unrealistically) reached a constraining boundary. This was the case for all starting points indicated by previous work, and several tolerance settings. The values given for v were found to be unsuitable since plotted distributions with those values were entirely different to kernel density estimates on the derived series. This was also the case when the active-set algorithm was used.

As a result, optimisations were run where v values were kept fixed at their initial estimates. This was more successful, with the optimisation converging for each zone when the tolerance was set to 10^{-3}. However, the ML parameter value sets were tested by plotting acf's for the ξ_t^2 series, revealing that temporal correlations were certainly still present. An alternative method was therefore tried – continuation of the grid-search method, exploring all local minima with a resolution of 0.01 for all parameters. This method has the significant advantage that no *a priori* assumptions about the distribution of the $\xi_t$were necessary. The parameter values obtained in this way seemed more realistic, and examination of acf plots for the squared series revealed that the method was indeed much more successful. Some small but statistically significant correlations remained however, but these were eliminated by a small change to the grid search algorithm.

Instead of using the sum of correlation absolute values for the range 1 – 50 hours as the loss function, several related alternatives were explored. The most successful of these was the sum of absolute values for correlations exceeding the Gaussian white noise boundary, for the range 1 – 25 hours. This set of searches proved very computationally expensive but delivered excellent results, an example of which can be seen in figure 7.14 below. Examination of plots such as this one, clearly demonstrating a lack or cross-correlations, made it clear that the choice of univariate models was justified. The final model parameters are shown in Appendix A3.

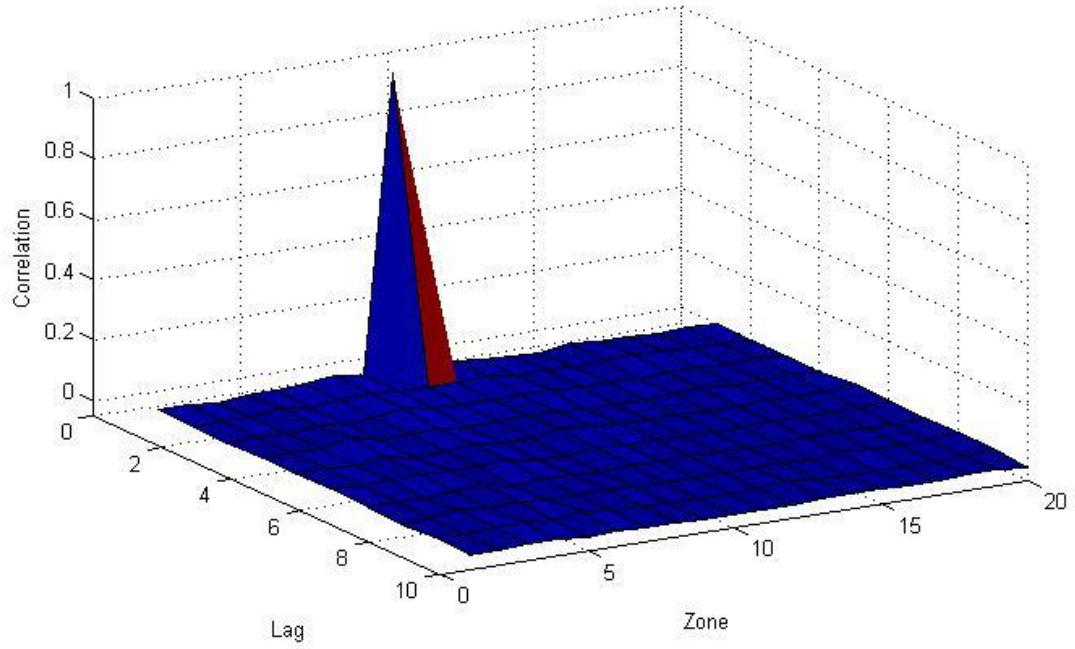

Figure 7.14. Auto- and cross-correlations between the squared i.i.d. residuals series at zone 9 and other zones, for lags 0 – 10 hours.

An unexpected and rather unfortunate consequence of the variance removal process was observed. Figure 7.15 below shows the familiar set of auto- and cross-correlations for zone 9, here it is for the $\xi_t$series (not squared). While figure 7.8 seemed to show that the z_t series for the zone was entirely free of temporal structure, a small but significant autoregressive tail may be seen in figure 7.15. Somehow, temporal correlation has been introduced.

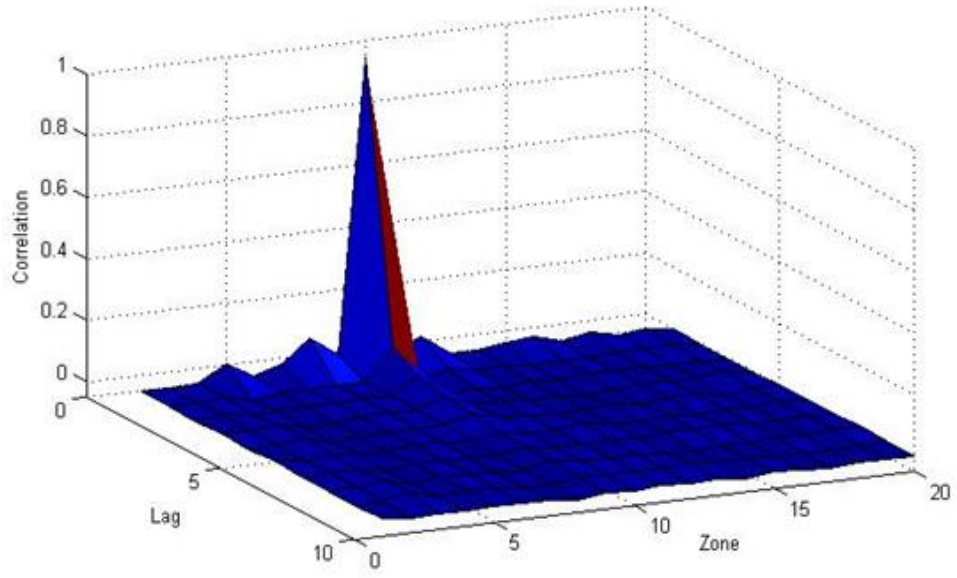

Figure 7.15. Auto- and cross-correlations between the i.i.d. residuals series at zone 9 and other zones, for lags 0 – 10 hours.

The only possible explanation is that there exist weak nonlinear temporally-lagged relationships between locations, which previously balanced out to leave zero linear correlation. However, the asymmetry of the variance model must have disturbed this balance, introducing correlation. Although unfortunate, this effect seems unavoidable, since the elimination of all such nonlinear relationships would be extremely challenging, and beyond the scope of this project. It is also worth noting that temporal correlations might not be introduced into the z_t series if obtained from simulated ξ_t - with accurate marginal distributions and linear spatial relationships, multiplied by the h_t generated by the model.

7.4. Chapter Summary

Given the results of the literature reviews and analysis of the previous chapter, the decision was made that the best initial choice for modelling wind expectation values was to assume that they form a vector Gagenbauer process, with annual seasonalities. The initial work was to find first estimates for the differencing parameters δ_i for each series individually. This involved selecting trial values for the δ_i then fractionally differencing the i^{th} series consistent with them, using the Gagenbauer polynomial series expansion for the back-shift operator, truncating at the 100,000th term. Once differenced, univariate AR models of different orders were fitted to the series using the Levinson-Durbin algorithm, and the lowest BIC value found. The resulting estimates indicated that long-memory is light, ranging from δ = 0.1 to 0.16. Algorithms were then developed, based on least-squares, to fit a full VGARMA model with optimal orders $p = 3$ and $q = 2$.

The next required stage was to attempt to refine the parameter estimates through MLE, clearly a challenge given the rather esoteric nature of the model and the high dimensionality. In order to facilitate this process, a thorough review was conducted on attempts by other authors to conduct exact, approximate and quasi- MLE on related models. It was clear from the review that exact maximum likelihood would be infeasible in this case, rather that a quasi-maximum likelihood estimator developed by Whittle should be used. This estimator makes use of the theoretical spectral density for a proposed model and empirical periodogram for the sample. Despite considerable effort, it was established that consistent Whittle estimates for the parameters could not be found for models with dimension greater than one.

One striking feature discovered when comparing theoretical and empirical spectra for each zone was that the empirical annual-period spikes are much bigger than their theoretical counterparts. This could be partially the result of rounding errors in the denominator for the latter, but could also be explained by the annual seasonality consisting of both deterministic and stochastic components. It was found that removing deterministic annual trends, carefully chosen to be optimally smoothed, removed the spikes but left the rest of the periodograms nearly identical. Further, the new periodograms were still consistent with long memory processes, with the long memory presence light.

The observed spectra could now be characterised as consisting of up to 100 spikes at the low frequency extreme. Several of these spikes were relatively large, in the range of about 1/3 to 1/2 of the height of the OLS-fitted model's spectral peak. Following heavy smoothing of

the periodograms, a universal pattern was found: a clump of the highest intensity centred between 1 and about 100 Fourier frequencies, along with several smaller clusters – one of which is slightly more prominent and centred between 60 to 400 Fourier frequencies. It therefore seemed more accurate to assume that there are two Gagenbauer frequencies. The zonal Gagenbauer frequencies were established by visual examination of the periodograms smoothed to various extents. With regard to the differencing parameters, it was found for a sample of zones that BIC values were reasonably close to minimal when the 2nd differencing parameter was 0.03, and it decided that the model will adopt this parameter value for each zone.

Algorithms were developed for the OLS-based fitting of a full multivariate model given the constraints above, with the optimum structure for the short-memory aspect emerging as ARMA(3,1). Attempts to improve upon the parameter estimates using the Whittle method proved unsuccessful once again. Estimates for the free differencing parameter show considerable range for different zones, varying from 0.03 to 0.13. It may be the case however that the presence of long memory does not differ quite this much, as there appears to be a type of colinieality between this parameter and the corresponding diagonal element in the 1st AR coefficient matrix.

It was confirmed that errors from the fitting of linear regression models are heteroskedastic, even after power-transformation and removing seasonality in the mean. It was shown initially that the VGARMA model is very successful in that no significant temporal correlations exist between errors, but that this is not the case for squared errors. Fortunately, correlations are largely limited to being auto-regressive in nature such that univariate GARCH-type models may be reasonably applied, avoiding the extreme complexity of multivariate models of this type. Volatility clustering may be easily seen in time series plots. It is also shown that patterns exist over the course of the annual cycle for the mean value of error variance, along with the inter-annual variability of error variance, and that these patterns are almost certainly not sample noise.

Removing this seasonality in variance was inherently a more challenging task than the removal of seasonality in mean, since it had to be achieved through division and power transformation alone, rather than subtraction and division. In particular, it was found that naïve approaches tend to over-compensate i.e. create new seasonal profiles that are roughly mirror images of the original pattern. However, a successful algorithm was developed, and described in this chapter. An assumption is made, without investigation, that the removal of

deterministic seasonalities in both the conditional mean and noise aspects of the wind speed process adequately accounts for the fairly weak positive correlation between mean and variance observed in Chapter 5. In other words, it is assumed that there is no need for an ARCH-in-mean element of the model structure.

The next and final part of the model building process was to fit univariate GARCH-type models to the de-seasonalised residuals. It was decided that the APARCH structure is most suitable since time series segments gave cause to believe that significant asymmetrical effects might be present. It was judged that autocorrelation functions decayed rapidly enough to set orders p and q as 1, and given the already very large number of fitted parameters it was decided to simplify matters by fixing the power parameter $\delta = 2$.

It was hoped that model parameter estimates could be established through maximum likelihood optimisation. This requires an assumed probability distribution type for the residuals, which are heavy tailed. Two alternative distributions were available, student-t and GED, since the relevant likelihood functions are given in the literature review. Empirical distributions for the residuals were estimated through kernel density estimation, and compared with best fitting parametric curves – finding that GED is a better for most zones. Unfortunately such best fitting value for the original, conditional errors were the only initial estimates available for the unconditional errors, despite a probable difference in their leptokurtosis. It was found, once again, that maximum likelihood estimation was not successful. The values given for the distribution parameter v were found to be unsuitable, with estimated distributions entirely different to kernel density estimates on the derived unconditional series. Assuming the unconditional errors to have the same parameter value as the estimated conditional ones was not succesfull either, since errors normalised according to the estimated values were not free of heteroskedasticity.

However, a (computationally expensive) grid search method was developed that was able to remove heteroskedasticity very successfully. Several cost functions were explored for this optimisation, the best one found to be the sum of absolute values for temporal correlations exceeding the Gaussian white noise boundary, for the lag range 1 – 25 hours. Some small correlations were unfortunately introduced to the non-squared unconditional errors, presumable arising from some nonlinearities that were previously hidden.

Chapter 8

Simulation Methodology and Validation

This Chapter presents the development of algorithms allowing accurate simulation of the random process described by the fitted model. This process was developed in reverse order to the model fitting, i.e. starting with i.i.d. noise and adding layer upon layer of structure until finally wind speeds were produced. During this process, analysis of results indicated that some changes and additions to the model were necessary, most notably the development of a transition matrix model.

The chapter also presents the methodology involved in the conversion of wind speeds to zonal power outputs for an example distribution of wind capacity, for the historical and synthetic series. This allows multi-faceted validation of the model.

8.1. Generating the Synthetic Series

8.1.1. Generating Correlated Deviates from Non-Parametric Distributions

The first simulation challenge was to develop an algorithm capable of generating temporally independent random deviates that have the same distributions as the derived ξ_t series from the previous section, both the marginal distributions for each zone, and their joint distribution. It was reported in the previous chapter that the g.e.d. parametric family was able to provide a reasonably good fit for each marginal distribution. However, this approach was abandoned since (i) no Matlab toolbox has a function for generating deviates from a g.e.d. distribution, and (ii) there are no specific multivariate extensions to such distributions.

There are several procedures that one may use to generate a set of deviates for random variables with a given joint distribution. One may use copula functions since, by virtue of Sklar's theorem, one need only to generate a set of $U(0,1)$ random variables whose joint distribution function is the copula function for the original random variables of interest. Then, the uniform random numbers for the i^{th} zone can be transformed to have the correct marginal c.d.f. $F_i(\xi)$ through it's quasi-inverse, i.e. $\xi = F(u)_i^{-1}$. However, this leaves the very substantial problem of specifying the 20-dimensional copula function.

The problem of fitting general multivariate copula functions may be simplified through factorisation of the copula function into a vine structure of pairwise copulas, as described for example in [137]. Despite being simpler, in such structures all pair copulas are conditioned on all of the remaining variables – up to 18 for the current problem. Fortunately, since spatial relationships are weak, one can make the drastically simplifying assumption that the form of the pair-copulae do not depend on these conditioning variables. Rather, it is assumed that dependency is purely through the distribution functions that constitute their arguments. Even with this simplification however, there remains a very large number of copula functions to be fitted, and a multitude of candidate parametric families. This makes the assumption that all relationships are Gaussian, i.e. are described by Gaussian copulas very appealing.

Even with such simplifications, the process of generating 20-dimensional deviates in this way remains a complex and expensive one. It was therefore decided that a very crude and therefore simple but mathematically dubious approach would be tried – and the approach proved surprisingly successful. The first step of this approach was to estimate the set of $F_i(\xi)$ (and consequently $F_i^{-1}(u)$) through re-arranging the series in increasing order – i.e. the largest negative values first. Using these, the series were transformed to being U(0,1) distributed and

the covariance matrix Γ_U for the transformed series calculated. Simulation of $\xi_t$series then consisted of:

- Generating a series of 20-dimensional vectors, whose components are all independent $U(0,1)$ deviates.
- Introducing spatial correlations by pre-multiplying the vectors with the lower-triangular Cholsesky decomposition of Σ_U. The lower rather than upper triangle was chosen since it will have the least effect on the generally more important northern-most zones.
- Re-establishing the boundaries of the marginal distributions. Since uniform distributions are not preserved under linear transformations, the correlated marginal distributions became rounded at their edges, with some deviates 'spilling over' the lines at 0 and 1. This effect was rather mild, since off-diagonal elements in the matrix are small. This was rectified with an algorithm that took the offending deviates and re-appropriated a value drawn from a normal distribution to them. Specifically, deviates less than zero were changed to deviates from $|N(0,\sigma^2)|$, with $\sigma \ll 1$, while deviates > 1 were changed to deviates from $1-|N(0,\sigma^2)|$. Several candidate values for σ were tried before finding one for each zone that left the marginal distribution looking exactly like U(0,1).
- Transforming the correlated and approximately U(0,1) distributed series to having the same marginal distributions as the historical ξ_t series by transforming according to $\xi_{t,i} = F_i(u_{i,t})$.
- Re-scale to ensure the new distribution has exactly the same standard deviation as the sample.

The success of this method was evident from kernel density estimate plots of the historic and simulated marginal pdf's – they were nearly identical for each zone. Also, the covariance matrix was calculated for the simulated ξ_t series and found to be in excellent agreement with the matrix for the historical series.

8.1.2. Generating the Heteroskedastic Noise Series

It was presumed that having generated the ξ_t series, generating the de-seasonalised z_t would be a trivial matter. The simple first algorithm, applied to a 100,000 hour set for each zone, was to assume that at $t = 0$, $h_t = \overline{h_t}$ and $z_t = 0$, then move along the series calculating h_t from z_{t-1} and h_{t-1}, then z_t from ξ_t and h_t. However it was found that, for each zone, at some early point in the series the variance exploded – growing until the largest number

Matlab can store was reached. Investigation revealed that, due to a positive feedback inherent in the algorithm, at some random point in each series a cluster of extreme values for z_t occurred close enough to each other to trigger the explosive instability.

Initial thoughts were that parameter constraints required some adjustment to allow for the heavy-tailed nature of the distributions. To investigate, model parameters were arbitrarily adjusted and the algorithm re-applied. Initially, α_1 was gradually reduced with β_1 kept constant and α_0 increased accordingly. For each zone, this was followed by β_1 being decreased with α_1 constant and finally both being decreased simultaneously. It was found that none of these changes solved the problem – at some point the variance always exploded. Decreasing the parameter values did however make the threshold for instability more extreme, such that the explosion generally occurred much further into the series.

It may be the case that if the power parameters δ in the APARCH model structures were not kept fixed in the model fitting stage, values considerably lower than 2 would have been found, possibly below 1 even, and this problem might not occur in that case. However the computational burden of such grid-searches is infeasible.

Another approach tried was to set a maximum value for h_t, with several values tried for several zones. Unfortunately, it was found that this approach lead either to extended periods where h_t hovered close to this limit, or to excessively long periods with quite distinct regimes of high and low variance. Other 'fixes' were tried which attempted to nudge large variances back towards smaller values, but none were successful in recreating realistic dynamics. It was therefore clear that the root of the problem was the aspect of positive feedback between h_t and $|z_{t-1}|^2$, and that this feedback somehow had to be removed.

The solution was to replace the defining equation

$$h_t = \alpha_0 + \alpha_1(|z_{t-1}| - \lambda z_{t-1})^2 + \beta_1 h_{t-1} = \alpha_0 + h_{t-1}(\alpha_1(|\xi_{t-1}| - \lambda\xi_{t-1})^2 + \beta_1) \qquad (8.1)$$

with one involving a new random variable Υ_t with values ν_t:

$$h_t = \alpha_0 + \alpha_1\nu_t(|\xi_{t-1}| - \lambda\xi_{t-1})^2 + \beta_1 h_{t-1}. \qquad (8.2)$$

The hope was that this definition can work as long as an algorithm could be developed that generates ν_t values with accurate conditional distributions $F_i(\nu|\Xi = \xi)$ for each zone i, as found in the historical sample. Such an algorithm proved elusive, but after many attempts a successful one was developed.

An aspect of the algorithm is that one must work with a transformation of ν, given by

$$\kappa = \sqrt{\ln(\nu)} \text{ for } \nu \geq 1, \quad \kappa = -\sqrt{|\ln(\nu)|} \text{ for } \nu < 1. \tag{8.3}$$

It also requires estimation of the marginal distributions $F_i(|\xi|)$ and $F_i(\kappa)$ through their derived values for the historical series, arranged in increasing order. The methodology divides the series into 20 bins of 5 percentiles for $F_i(|\xi|)$, before calculating the conditional probability distributions $F_i(\kappa||\xi|)$. Whilst doing this for narrower bins of e.g. 1 percentile would have provided a more refined picture of the effect of conditioning, the sample size within each bin would have been smaller and the results noisier.

Having extracted all necessary information from the historical sample, generation of de-seasonalised heteroskedastic noise was achieved as follows:

- Generate a matrix of 20 ξ_t series, as described in section 8.1;
- For each matrix element $\xi_{i,t}$ calculate $F_i(|\xi_{i,t}|)$ and the bin in which it lies;
- Generate an $U(0,1)$ random number and use it, along with knowledge of probabilities $F_i(\kappa||\xi|)$ for each bin, to set the $F_i(\kappa_{i,t})$ bin. Generate another uniform random number to give $F_i(\kappa_{i,t})$ a precise value within the bin;
- Convert to $\kappa_{i,t}$ and then to $\nu_{i,t} = \exp(\kappa_{i,t}.|\kappa_{i,t}|)$;
- Calculate $h_{i,t}$ according to equation 8.2 and consequently$z_{i,t}$; and
- Re-scale by calculating $\overline{|z_t|}$ for the historical and synthetic series, and multiply the $z_{i,t}$ by the factors $(\overline{|z_t|}_i\ historical)/\ (\overline{|z_t|}_i\ synthetic)$.

It was found that the algorithm produces some extreme values that are more than six times greater than any found in the historical series, even after the re-scaling step above. Although it is entirely natural for the synthetic series to contain tail values that are more extreme than those in the historical series, the extent of the difference here is almost certainly a result of a shortcoming in the model or algorithm. In order to establish a reasonable cut-off, we return to considerations of the definition of long memory in Chapter 2. Here we slightly re-write equation 2.8, with the length of time T replaced by the sample size n:

$$E[R/S] = Cn^H, \text{ with } H > 0.5. \tag{8.4}$$

Although $H > 0.5$ for long memory processes, for uncorrelated processes without long memory, such as z_t, $H = 0.5$. Therefore for two sample sizes n_1 and n_2, if we assume the standard deviation to be independent of sample size, we get $E(R_2/R_1) = \sqrt{n_2/n_1}$. Since $n_2 = 100{,}000$ for our synthetic sample and $n_1 = 55{,}287$ for the historical sample, this gives $E(R_2/R_1) = 1.34$. Allowing for the fact that this ratio is merely an expectation value, it was decided that for historical series i if we label $z\_max_i = \max(|z_{i,t}|)$, then the magnitude of

extreme values in the synthetic series should not be allowed to exceed $1.4 * z\_max_i$. An addition to the noise generation algorithm was therefore developed which takes excessively extreme values from the series and redistributes them randomly and uniformly in the ranges $-z\_max_i$ to $-0.6 * z\_max_i$ or $0.6 * z\_max_i$ to z_max_i, as appropriate.

It was found that one final re-scaling led to slightly improved fits. This involved calculation of the maxima of the (Gaussian) kernel density estimates (KDE) for the historical and synthetic series, labelled $kde\_max\_h_i$ and $kde\_max\_s_i$ respectively, then multiplication of the $z_{i,t}$ by the factor $kde\_max\_s_i \; / kde\_max\_h_i$. Following such re-scaling and trimming, KDE plots for the historical and synthetic series displayed excellent agreement, as exemplified by figure 8.1 below.

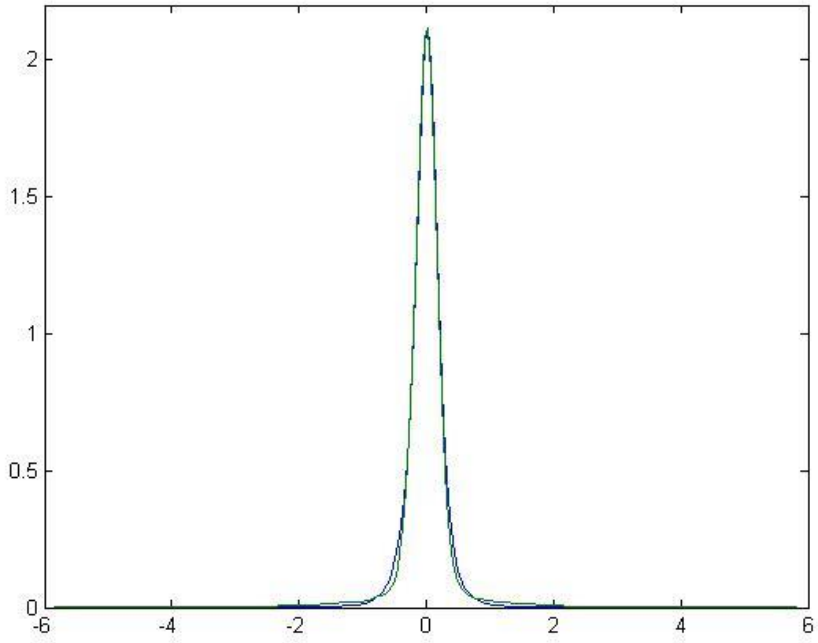

Figure 8.1. Kernel density estimates for synthetic and historical de-seasonalised noise series, zone 4

The next stage in the simulation was to reinstate seasonality, both intra-annual and inter-annual variability profiles. This was a straightforward task, simply a reversal of the algorithm which removed these patterns. A re-scaling was necessary, again of the form $(\overline{|z_t|}_i \; historical) / \; (\overline{|z_t|}_i \; synthetic)$, in order to ensure that synthetic and historical variances match. Plots of synthetic and historical kernel density estimates with seasonality reinstated showed the same excellent agreement.

8.1.3. Constructing Wind Speeds from the Synthetic Noise

The defining equation for the process being simulated is

$$\boldsymbol{\Phi}(B)\,\underline{U_t} = \boldsymbol{\Theta}(B)\underline{z_t}, \tag{8.5}$$

where $\underline{U_t}$ is the vector process of the twice fractionally differenced transformed wind speeds, the AR matrix polynomial $\boldsymbol{\Phi}(B)$ is 3rd order and the MA matrix polynomial $\boldsymbol{\Theta}(B)$ is 1st order. This may be expanded and re-arranged to give

$$\underline{U_t} = \boldsymbol{\Phi_1}\underline{U_{t-1}} + \boldsymbol{\Phi_2}\underline{U_{t-2}} + \boldsymbol{\Phi_3}\underline{U_{t-3}} + \underline{z_t} + \boldsymbol{\Theta_1}\underline{z_{t-1}}, \tag{8.6}$$

so that the doubly-differenced series may be constructed, starting from t = 1, with values from the historical series used for $t \in [-2,0]$. This was done for a 10-year sample of $\underline{z_t}$.

With the doubly-differenced series generated, the next task is to reverse the differencing, i.e. integrate. For this we introduce the variables $V_{i,t}$, representing the once-differenced series $\underline{V_t}$ at zone i. Here we may work in univariate terms, since the differencing matrices are diagonal. Recalling equation 3.30 from section 3.2.1 describing the expansion of the differencing operator, we therefore have that

$$(1 - 2v_{1,i}B + B^2)^{-\delta_{1,i}}\, V_{i.t} = \left(\sum_{k\geq 0} C_k(\delta_{1,i}, v_{1,i})\, B^k\right) V_{i,t} = U_{i,t}, \text{ and} \tag{8.7}$$

$$(1 - 2v_{2,i}B + B^2)^{-\delta_{2,i}}\, X_{i.t} = \left(\sum_{k\geq 0} C_k(\delta_{2,i}, v_{2,i})\, B^k\right) X_{i,t} = V_{i,t}, \tag{8.8}$$

where the C_k are Gagenbauer polynomial coefficients. Further recalling that $C_0(\delta, v) = 1$, and truncating the expansions at $k = M_1$ and $k = M_2$ respectively, we may re-write these equations as

$$V_{i,t} = U_{i,t} - \sum_{k=1}^{M_1} C_k(\delta_{1,i}, v_{1,i})\, V_{i,t-k}, \text{ and} \tag{8.9}$$

$$X_{i,t} = V_{i,t} - \sum_{k=1}^{M_2} C_k(\delta_{1,i}, v_{1,i})\; X_{i,t-k}. \tag{8.10}$$

To generate a synthetic series for $X_{i,t}$ of length N we may therefore start at $t = 1$ and use synthetic values of $U_{i,t}$ for $1 \leq t \leq N$, and a segment of historical $V_{i,t}$ series for $-M_1 < t < 1$ to generate a synthetic series of $V_{i,t}$, before similarly using this series along with a historical segment of $X_{i,t}$ to generate the synthetic $X_{i,t}$ series. Initially, both M_1 and M_2 were set as 25,000 so that several cycles of each quasi-frequency were included.

Despite being significantly shorter than the previous truncations (70,000 and 50,000 respectively) the integration process was still computationally expensive, and thus only a 10-year series was generated. Following integration, deterministic seasonalities were re-introduced: first the annual seasonality in mean, then diurnal seasonalities in variance and mean.

8.1.4. Final Adjustments and Censoring of the Synthetic Series

It was found that the means of the synthetic series, at this stage of reconstruction, were not exactly zero. Deviations were not very large, but significant enough to be suspected as 'unrealistic'. A definite answer to that question was not possible without historical series of annual mean wind speeds stretching back at least 100 years. However in order to gain some insight, 10-year means were calculated for the 11 different 10-year segments contained within the equivalent 20-year historical series, for each zone, and the largest deviations from zero recorded.

Whereas the annual means of the historic (transformed) series have standard deviations ranging from 0.074 for zone 1 to 0.165 for zone 20, the largest deviations from zero of a historic 10-year mean range from 0.018 for zone 1 to 0.12 for zone 2. The means of the synthetic series were roughly 5 – 6 times greater, which is clearly unrealistic. It was therefore decided that the deviation of the synthetic series at this point should be capped at 20% greater than the maximum deviation in the historical series, keeping the direction of deviation. The standard deviations of the synthetic series were also checked at this point, and encouragingly were found to deviate from the equivalent historical series by no more than about 10% in almost cases.

However, kernel density estimates revealed that the distributions of the synthetic series were all more leptokurtic than their historical counterparts. Various algorithms were therefore developed with the goal of transforming the synthetic distributions to closely resemble the historical ones, based on power transformations. The best approach found was to find a number close to 1 for each zone, ι_i, then transform the synthetic series according to

$$\begin{aligned} x_{i,t} &\rightarrow \iota_i^{(1-|x_{i,t}|)}\, x_{i,t} \quad \text{for} \quad |x_{i,t}| \leq 1 \\ x_{i,t} &\rightarrow x_{i,t} \quad \text{for} \quad |x_{i,t}| \leq 1. \end{aligned} \tag{8.11}$$

The optimal constants ι_i were established by grid search and visual inspection of kernel estimated densities – with both the newly transformed series and equivalent historical series plotted. Optimal values were found to be > 1 for all but one zone, and ranged from 0.95 – 1.15. The transformation therefore has the effect of increasing the relative absolute value of all $|x_{i,t}| \leq 1$, with the effect most pronounced for the smallest values. This made the central peak of the distributions more rounded, as desired. Following the power transformation, division was used to re-scale the series to have the same variance as the equivalent historical series.

The last simple stages of the wind speed reconstruction process were then applied – adding long-term means, making all negative values zero, raising all values to the power 2.5 and rounding to the nearest knot.

8.2. Initial Analysis of the Synthetic Series

8.2.1. Wind Speed Distributions at Single Locations

With a ten year sample generated, initial analysis of the series' similarity to the historical series was possible, as reflected through their distribution. Comparisons of long-term values for mean, median, mode, standard deviation and skewness were made for each zone.

It was found means are in good agreement, although synthetic series means are slightly larger in each case. The difference is less than 5% of the historical means for most zones, and the largest difference is 13%, for zone 10. Medians were found to be in excellent agreement, with the majority identical (in integer knots), while none were greater than 1 knot different. Modes were mostly in fairly good agreement, typically differing by about 2 knots – except for the minority of zones where the historical series mode is zero knots, or close to it. In this respect, the synthetic series are probably the most realistic. Standard deviations are in good agreement, larger for synthetic series in each case but mainly within 15%, the largest difference being 26% for zone 10. Skewness is the least similar of measures – while this is typically about 0.8 for the historical series, the synthetic series have values centred arround about 1.25 and zone 9 has more than twice the historical value. However, the ratio is almost exactly unity for 3 zones and is less than 1.5 for an additional 8 zones. Despite the kurtosis-matching algorithm described above, the final synthetic series were universally more leptokurtic than their historical counterparts. Information on moments for three example zones is tabulated in Appendix 4, as are distribution percentiles for those zones.

The degree of similarity between the series seems greater when looking at their histograms, which were plotted as line graphs for each zone, and taken to be their approximate marginal distributions. Figure 8.2 shows a slightly better than average success – zone 3, while figure 8.3 shows zone 6, where the historical series deviates very strongly from the Weibull distribution.

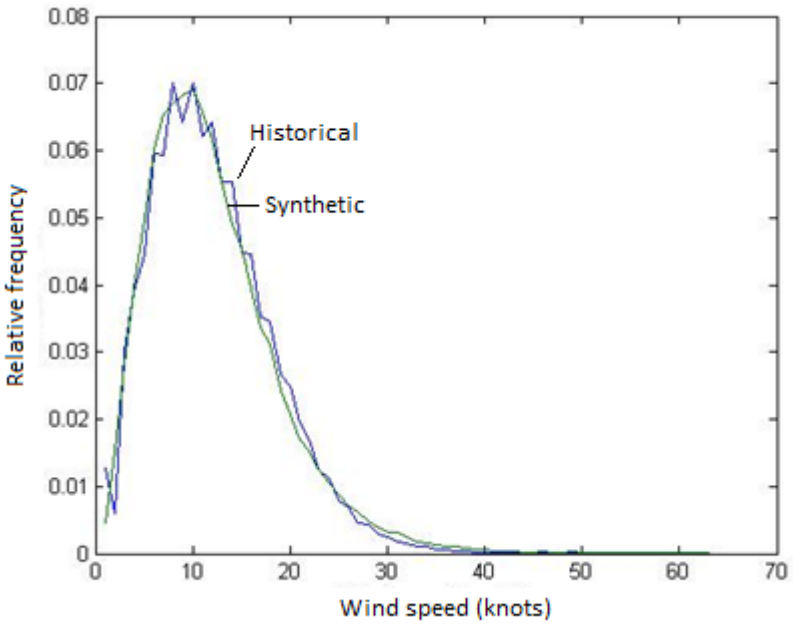

Figure 8.2. Approximate marginal distribution of wind speeds, historical and synthetic series, zone 3.

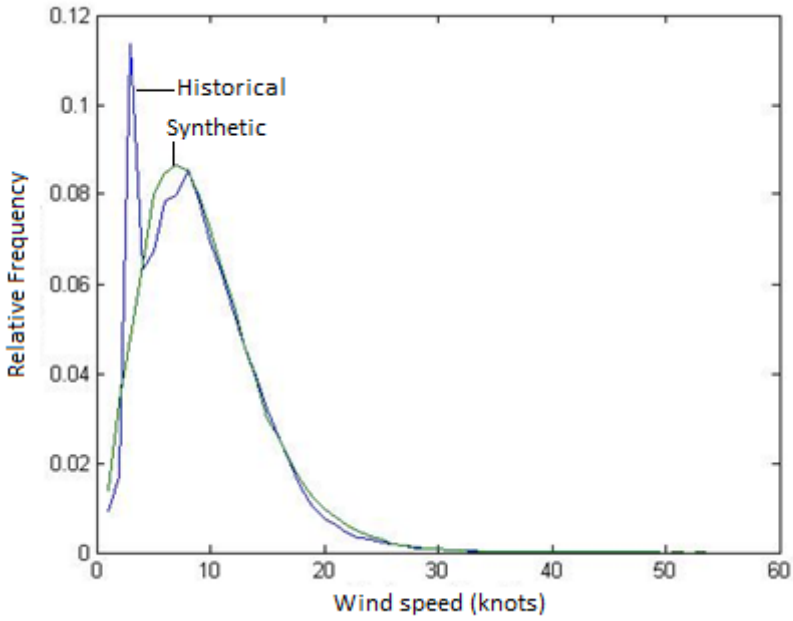

Figure 8.3. Approximate marginal distribution of wind speeds, historical and synthetic series, zone 6.

8.2.2. Wind Power Distributions – Single Locations and Aggregated

Having compared synthetic and historical distributions for 10m wind speeds, it is also informative to also compare implied power outputs – with the very important caveat that the methodology of converting wind speed to power is rather crude, and the results should be considered purely illustrative.

The initial stage is to 'up-scale' the wind speeds, by first accounting for superior resources at real wind-farm locations, and then hub-height increases. The simple approach adopted was to assume fixed typical values for these, obtained from up-scaling tables found in the report prepared for National Grid [27], described in Chapter 1. They are a 20% increase for location and a 30% increase due to height, for all zones. The next stage was to convert from speed to normalised power using a fixed, wind-farm scale effective power curve, derived using a combination of power curve tables found in [27] and typical farm-scale spatial smoothing effects reported in [147]. These were further multiplied by the assumed availability factor of 0.95, avoiding the sampling procedure adopted in [27] to account for stochastic turbine availabilities.

Following conversion of both synthetic and historical series, results were plotted for each zone. Since the wind series took only discreet values, i.e. integer knots, wind power outputs were also discrete, and it was the relative frequencies of these output levels that was plotted. It was found that synthetic and historical series were in very close agreement for all zones, particularly intermediate power levels. A typical example, zone 14, is shown in figure 8.4., where there is slightly better than average agreement at the edges. Another good fit, which is actually one of the least accurate, is shown in figure 8.5. – zone 13. Whereas the histograms are almost flat at intermediate values for most zones, the relatively poor resource at zone 13 means that there is a negative gradient over such values, which is reflected well in the synthetic series.

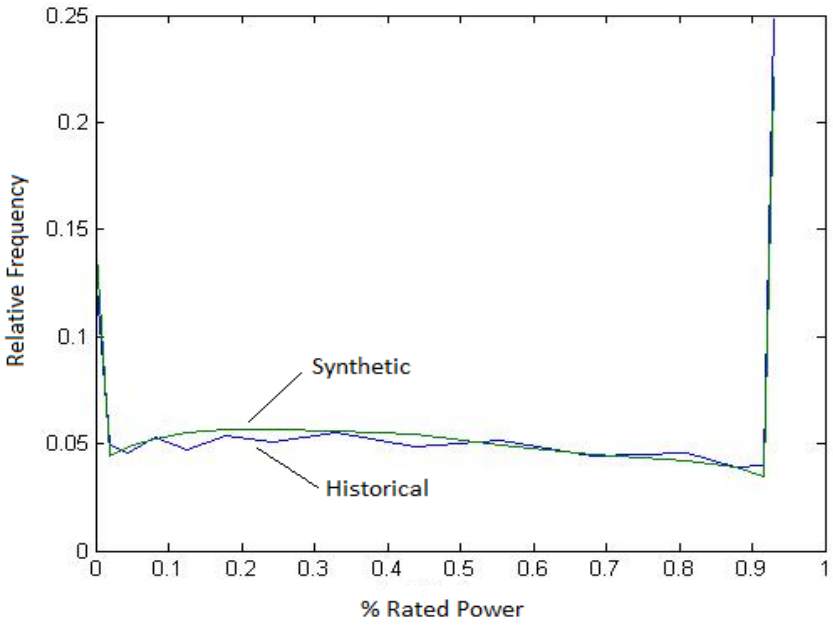

Figure 8.4. Approximate marginal distribution of power outputs, historical and synthetic series, zone 14.

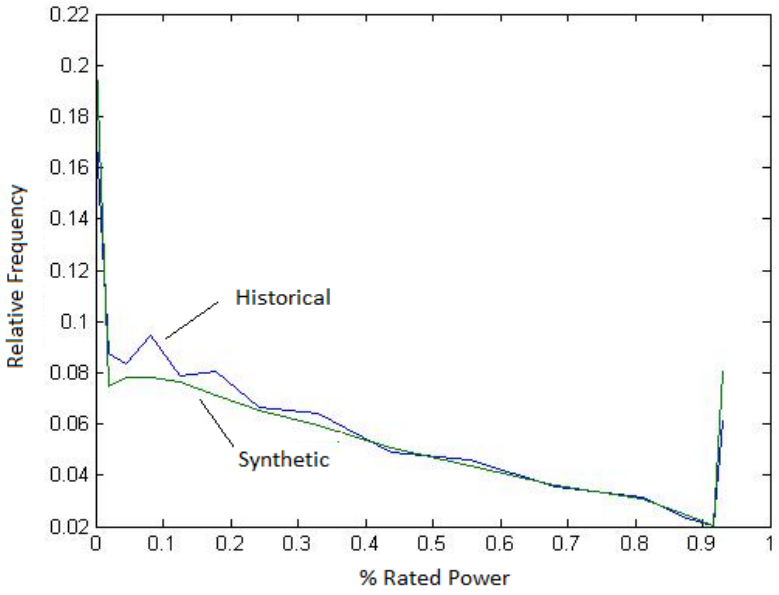

Figure 8.5. Approximate marginal distribution of power outputs, historical and synthetic series, zone 13.

The marginal power output series were added together, without regard to realistic zonal weighting, to produce a rough estimate of the aggregate output's distribution, with the results shown in figure 8.6.

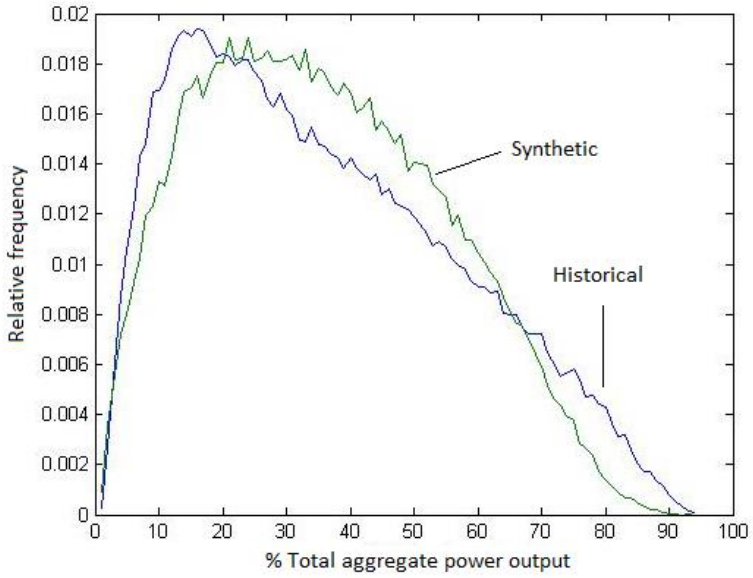

Figure 8.6. Approximate distribution of aggregated power outputs, historical and synthetic series.

The most notable feature is that, as expected, the two extremes of zero and full rated power outputs are no longer the most frequently occurring – rather, it is the lower end of the intermediate output levels. Such output levels represent highly varied combinations of individual zonal power outputs. Unfortunately, the extent of agreement between the historical and synthetic datasets is lower than for individual zones, with the synthetic series under-representing the lowest and highest aggregate power states, i.e. < 10% and > 70%. However their means are in very close agreement: 34.94% for the historic series and 34.83% for the synthetic; also their medians are close: 32 and 33 respectively, and their standard deviations are quite close: 21.12% and 18.76%. Their differences may be partially explained by errors in the marginal distributions, but much more so by the implied inadequacy of the assumed Gaussian copula for the joint distribution of wind speeds. It seems likely that universally low and high wind speeds are more common than for a MVN process.

A vital question at this point, due to the highly nonlinear nature of the power curve, is the extent to which the above observation is dependent on the up-scaling of the wind speeds before conversion. To test this, the up-scaling was changed from a factor of 1.56 to 1.3, and the new aggregate probability distribution plot is shown in figure 8.7 below.

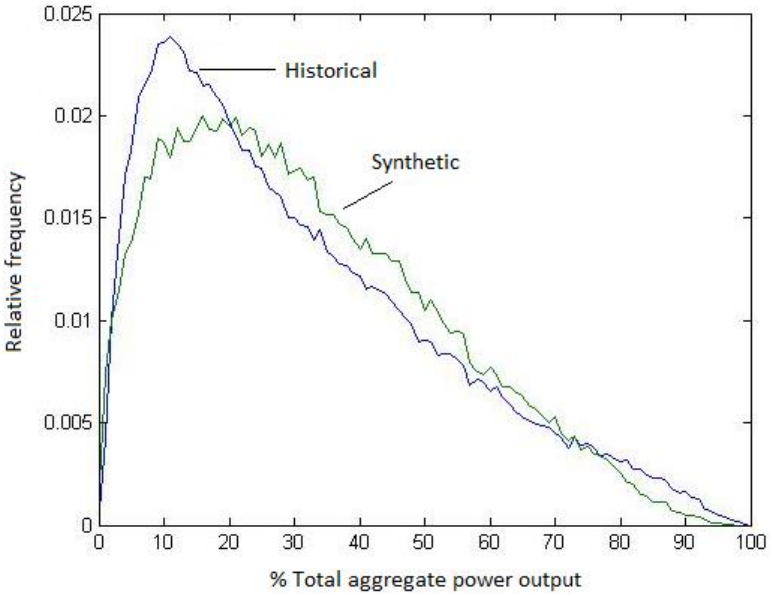

Figure 8.7. Approximate distribution of aggregated power outputs, historical and synthetic series, with reduced speed-up

The plot shows that indeed the same observation can be made – the major difference being that the low power under-representation is now worse, but better at the high power end. Most summary statistics remain in excellent agreement – the means being 26.16% for the historical series and 26.69% for the synthetic, for example. It is clear that a more precise examination of the distribution of vector states is needed, considering the original wind speeds rather than powers.

Another essential aspect to be replicated is the distribution of changes over a fixed period. Plots were prepared for the distribution of changes in the aggregate power, using the factor of 1.56 for up-scaling, for 1 hour and 4 hours. Both showed excellent agreement, with the 4 hour results shown in figure 8.8 below. This is very encouraging, since there was nothing in the model fitting that explicitly ensured that these would match.

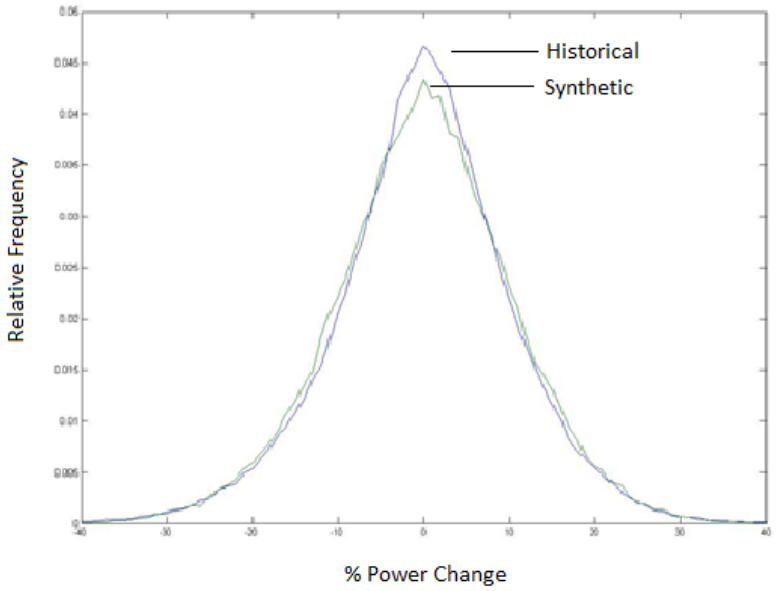

Figure 8.8. Approximate distribution of aggregated power output changes over 4 hours, historical and synthetic series

8.2.3. The Distribution of Vector Wind-Speed States

In order to compare the joint distribution of wind speeds in the synthetic and historic series, the wind speeds at individual locations were translated into discrete states. Then, the frequencies of occurrence of vector states, i.e. combinations of marginal states, were calculated. The approach was to work with wind speeds up-scaled by 1.56, and define 4 states, where the wind speed is: (i) lower than turbine cut-in; (ii) between cut-in and rated power; (iii) within the range corresponding to rated power; and (iv) above turbine cut-off speed. The set of GB wind states may then be defined as all possible combinations of individual states for the 20 zones. However, this represents 4^{20} states, i.e. 1.1×10^{12}, which is an impractically large number. The alternative is to merely distinguish the number of zones in each state, which is still larger than the practical limit of the sample length.

Each possible vector state was given an associated number, calculated according to (1 * Number of zones in state 2) + (100 * Number of zones in state 3) + (1000 * Number of zones in state 4). This means that e.g. when all zones are below cut-in speed the state is 0, and when all are above cut-off the state is 20,000. All states not found to occur in either the historic or synthetic series were then discounted and occurring states re-labelled in incremental steps of unity.

It was found that due to spatial correlation, the number of different vector states that occur in the historical sample is 216. Pleasingly, the number of vector states found in the synthetic series was 220, and the set of states covering both the historical and synthetic series has only 228 elements – i.e. there is very considerable overlap between the two sets. The vector states were therefore given labels from 1 to 228, and their relative frequency in the two sets is shown in figure 8.9 below. Whereas there is a clear trend of increasing windiness as the state index value increases, the states are not exactly ordered in terms of increasing power output. The 1st cluster of states clearly represent a situation in which there is no generation in some zones, and below rated capacity generation in all of the others. All clusters may be interpreted in a similar way, although interpretation is not trivial in many cases.

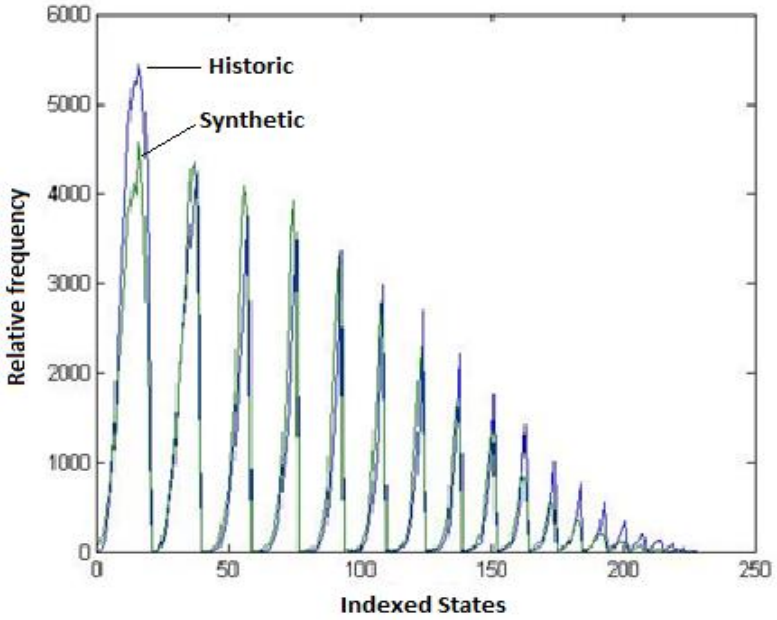

Figure 8.9. The distribution of indexed vector states in the synthetic and historical series

The plot shows that the two series are very similar, albeit with some noticeable flaws, which is a very important and positive result. Although it is difficult to interpret the indexed states directly, the plot seems to confirm that the synthetic model tends to under-represent states with a high degree of spatial similarity, and over-represent highly mixed states. In particular, there is an under-representation of the highest numbered states, i.e. those in which the entire country is very windy. In contrast to the general pattern, the synthetic model slightly

overestimates the occurrence of the 6 lowest states i.e. ones in which between 0 and 5 zones only are generating.

The question naturally arises: is there some way of correcting for these deviations from Gaussian copula behaviour, without having to explicitly model the high dimensional copula structure? If so, would this enhanced model lead to a noticeably improved aggregate power distribution? One possibility is to fit a multivariate, transition matrix based Markov chain model to a subset of zones, and this is the topic of the next section.

8.3 Enhancement with a Transition Matrix Model

This section describes the fitting and application of a multivariate transition matrix model in an attempt to make the spatially-joint distribution of the synthetic wind speeds an even better match to that of the historic series.

In this work, the meaning of 'multivariate' goes beyond its interpretation in Chapter 3, namely that the state of the process at zone i at time $t+1$ depends not only on its state at time t, but also on the state of (potentially all) other zones at time t. The limitation of such a model is that the state of the process at each zone is decided by sampling from $U(0,1)$ distributions that, in order to be realistic, should have a complex dependency structure captured by a non-Gaussian multivariate copula. Rather than simplifying by assuming the $U(0,1)$ deviates to be independent, the method adopted here is to make transition between vector states of the type described in the previous section. Since we are dealing with highly heteroskedastic processes, the vector states must be extended to include discretised conditional variance states as well as (transformed) wind speed states.

The basic concept is to choose a suitable subset of zones, then a suitable set of vector speed and volatility states for them, fit a transition matrix between the states and use it to simulate a series. As discussed in Chapter 3, a shortcoming of transition matrix models is that they under-represent temporal auto-correlations for lags beyond a fairly short range. For this reason, and also since they cannot reproduce seasonality, the transition matrix model must be fitted to the transformed and doubly-fractionally-differenced historical series at the chosen zones. The resulting simulated series will be free of the restrictions of the MVN assumption, and can potentially capture the joint distribution of the subset more accurately

Univariate ARMA-APARCH models should then be fitted to the same historical series, and applied to synthetic series generated by the transition matrix model to obtain series of unconditional N(0,1) residuals. Then, for each hour to be simulated, unconditional innovations may be generated for the remaining zones using the same correlation matrices and methodology as previously (section 8.1.1), but now also conditioning on the residuals obtained from the transition matrix model. The proceeding sections of methodology should then be followed to obtain the final wind speeds. The idea therefore is that the results of transition matrix model simulation may be fed into the simulation methodology for the entire set of zones, 'nudging' spatial relationships in the right direction when needed.

Clearly the biggest challenge in developing such a model is the very large number of vector states, even for a very coarse discretisation and small sub-group of zones. After exploring a number of options, it was decided to choose include 4 zones in the subset, each with 5 states for the wind speed and 4 states for conditional variance. This means that we are dealing with 8-dimensional vector states and potentially $5^4 * 4^4 = 160{,}000$ different states, although the number actually observed in the historical series will surely be considerably smaller. The choice of zones had to extend across the entirety of GB, without being too isolated from all neighbours, and those chosen were 3, 9, 16 and 17 (in north eastern Scotland, south western Scotland, eastern England and the south west of England, respectively). States were defined on the basis of equal probability, i.e. 20 percentile bins for the transformed wind speeds and 25 percentiles for the conditional noise variance.

The doubly differenced historical series is 55,287 hours in length. Within this sample 32,285 vector states (from a possible 160,000) were found, and these were taken to be the complete set of states in which the process may exist. Since the vector states are so specific, they all occurred very rarely. In fact, 65% of them occurred for one hour only, 20% of them for only 2 hours and 14% between 3 and 10 hours. This leaves less than 1% that occurred for more than 10 hours, and the most frequently occurring state did so for 52 hours. When the simulated series transitions into those states that occurred during only one hour, it must then transition with a probability of 1 into the same state as it did in the historical series – which is clearly unrealistic. Also, an artificial but plausible forced transition path had to be set up for the last observed state. Somewhat surprisingly, this flaw is far from fatal, and marginal distributions for the doubly differenced wind speeds were of a similar, excellent quality when generated with this model as when generated using the original method.

A comparison of the final wind speed distributions obtained with and without the transition matrix model is provided in Appendix A4, along with the same results for the historical period. Considering distribution quantiles, 28 of the 36 tabulated values were identical integers, the others being only one integer apart. Of the eight that are different, 6 were closer to the historical series' values when aided by the transition matrix model, which indicates that its inclusion has a very small positive effect. This contrasts with consideration of moments, since table A4.2 demonstrates that inclusion of the transition matrix model has somehow made the problem of excess skew and kurtosis worse. However, these errors relate to extreme values that simply correspond to zero output, so the slight improvements described above are probably more significant. It was also found that following conversion to

powers through the previously described procedure, the aggregate power distribution was very slightly closer to the historical one, as was the distribution of 4-hour changes. For these reasons, it is the transition matrix model enhanced series that will be used in the analysis that follows.

8.4 Analysis of the Synthetic Wind Speed Sample

This section reports on the results of a variety of experiments to further judge the extent of similarity between the synthetic and historical sample. For additional insight, a simpler model was also fitted and used for comparative purposes. This model kept the power transformation aspect, along with deterministic removal of seasonality in mean and variance on both diurnal and annual scales. However, long memory and conditional heteroskedasticity effects were ignored, the model fitted to the de-seasonalised series being VARMA, using OLS optimisation. The fully sophisticated model/ series generated from it will be referred to here as the Gagenbauer model/ series, while the simpler one will be referred to as the VAR model/ series.

Little additional insight would be gained from generating longer synthetic datasets than the 10 year one already discussed. There would hopefully be considerable benefits to generating very long datasets for use in Monte Carlo simulations – indeed this belief is the motivation for developing the time series models. However the natural superiority of very long series is negated here by the numerous censors and normalising transformations contained within the simulation algorithms that prevent the synthetic series from deviating 'too far' from the historical ones.

8.4.1 Temporal Correlations

Autocorrelation functions (ACFs) were calculated for the 3 series, for all zones, and the results plotted. The plots were for two ranges: small and medium sized lags of 0 – 500 hours and long lags of 1001 – 10,000 hours. It was found that in the 1st range, the Gagenbauer series generally replicated the correlation structure of the historical series quite well, and significantly better than the VAR series for most zones. In the long range, it was found that both synthetic model ACFs decayed more quickly than in the historic series, for all zones, and in many cases the two models were surprisingly similar. Figures 8.10 – 8.13 show these plots for the example zones 3 and 13, with the white noise correlation boundaries included.

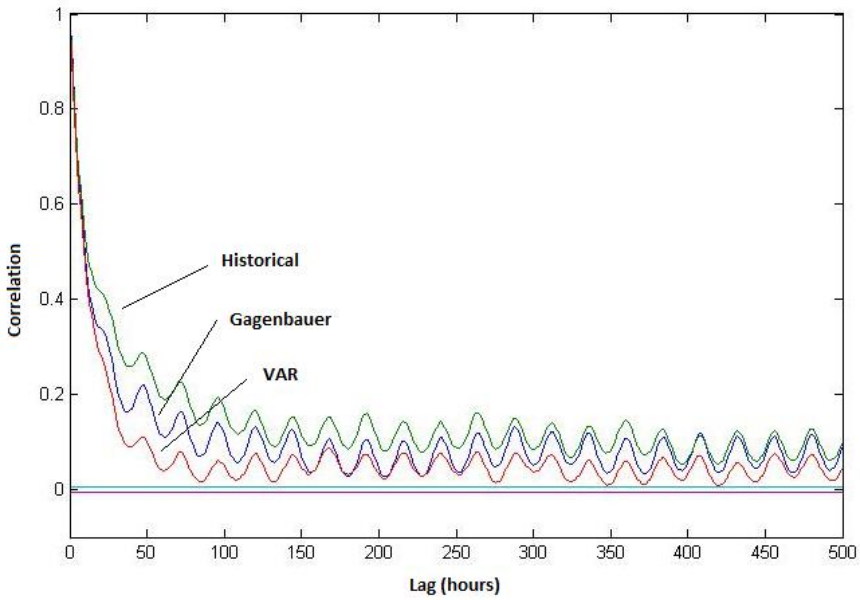

Figure 8.10. Short and medium range ACF's for the historical, Gagenbauer and VAR series

In the case of zone 3, it can be seen that the Gagenbauer series ACF is significantly closer to the historic one than the VAR ACF, but is particularly close from a lag of about 280 hours. For long lags, the Gagenbauer model is clearly a better match, although the artificial presence of the Gagenbauer frequencies as a jaggedness in that series.

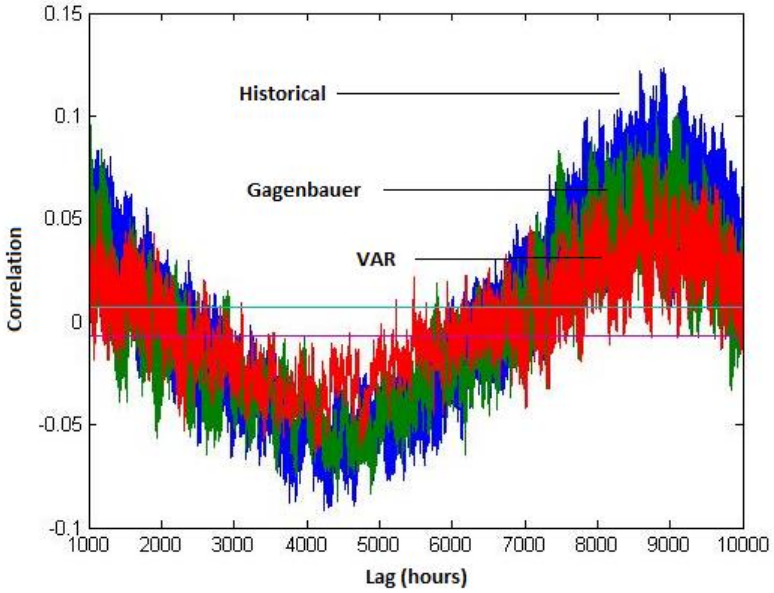

Figure 8.11. Long lag ACF's for the historical, Gagenbauer and VAR series, zone 3

In the case of zone 13, figure 8.12 shows that the Gagenbauer ACF is a much better fit for roughly the 1st 200 lags, but that for lags > 300 the AR model is a slightly better fit – surely an unfortunate consequence of the artificial Gagenbauer seasonality. Figure 8.12 shows that for lags > 1000 hours, both synthetic ACF's are roughly equally good fits, with the VAR model possibly superior as the curve is smoother. Such a poor outcome is fortunately atypical of most zones.

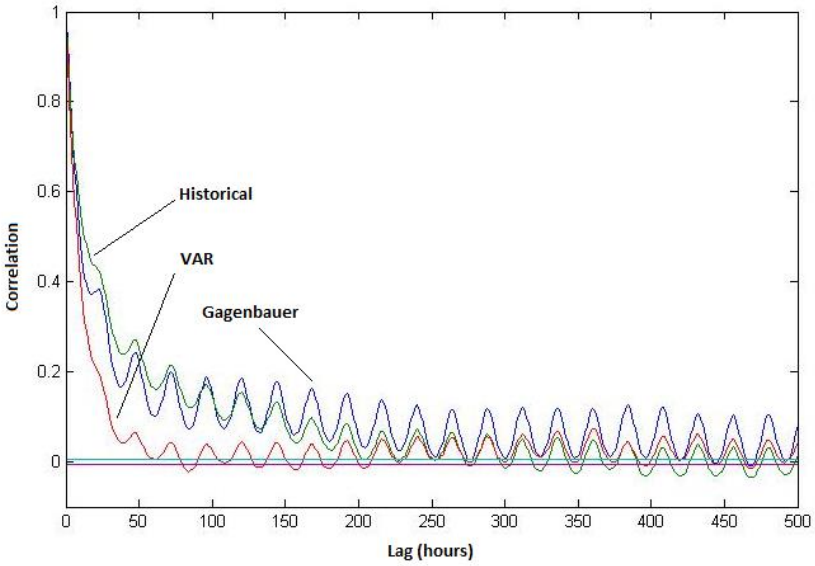

Figure 8.12. Short and medium range ACF's for the historical, Gagenbauer and VAR series

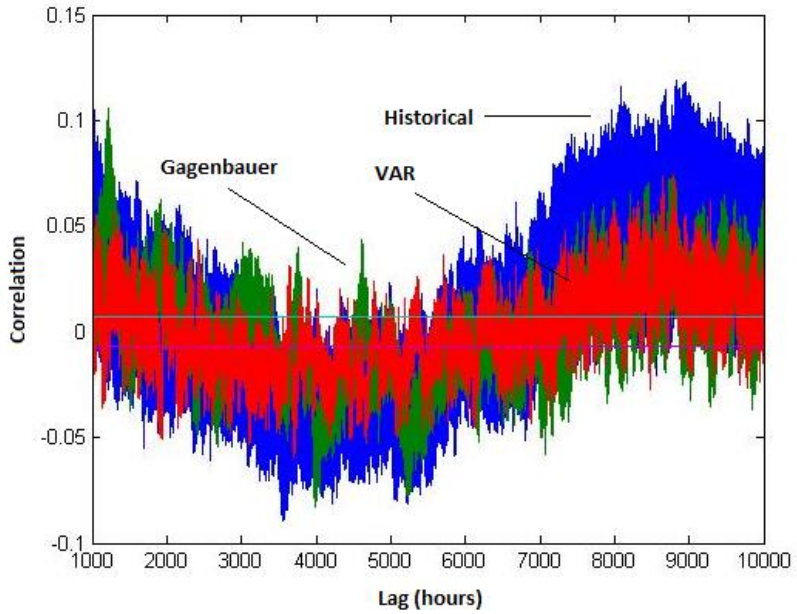

Figure 8.13. Long lag ACF's for the historical, Gagenbauer and VAR series, zone 13

8.4.2. Comparison of Spectra

Spectral intensities were calculated for the 3 series, for a selection of zones, and the periodograms plotted. In order to make the comparison equivalent, the calculations were carried out on the 10 year samples repeated (to form 20 year samples). The plots showed that both series were in good agreement, but the Gagenbauer series were generally better fits. Figure 8.14 shows the segment of periodogram surrounding the annual seasonality peak for zone 1, with the log Fourier frequency index on the horizontal axis. The match between the historical and Gagenbauer series is so good at the peak that they cannot be distinguished; the VAR model also being quite a good fit.

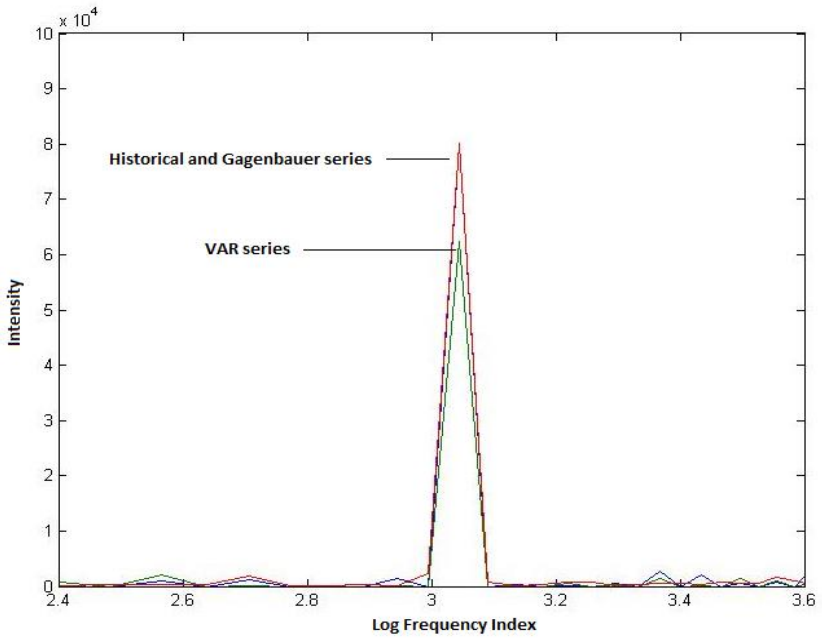

Figure 8.14. Periodogram segment surrounding annual peak, log frequency scale, zone 1

A similar segment is shown for zone 8 in figure 8.15, although the full range of sub-annual frequencies is displayed in this case. It shows that for this zone, at the annual frequency the Gagenbauer series is a good but not perfect match for the historical one, but the VAR model seriously underestimates the intensity. At sub-annual frequencies, none of the series are in agreement, which is not very surprising, but the Gagenbauer series at least has peaks of similar height to the historical series, while the VAR series certainly does not.

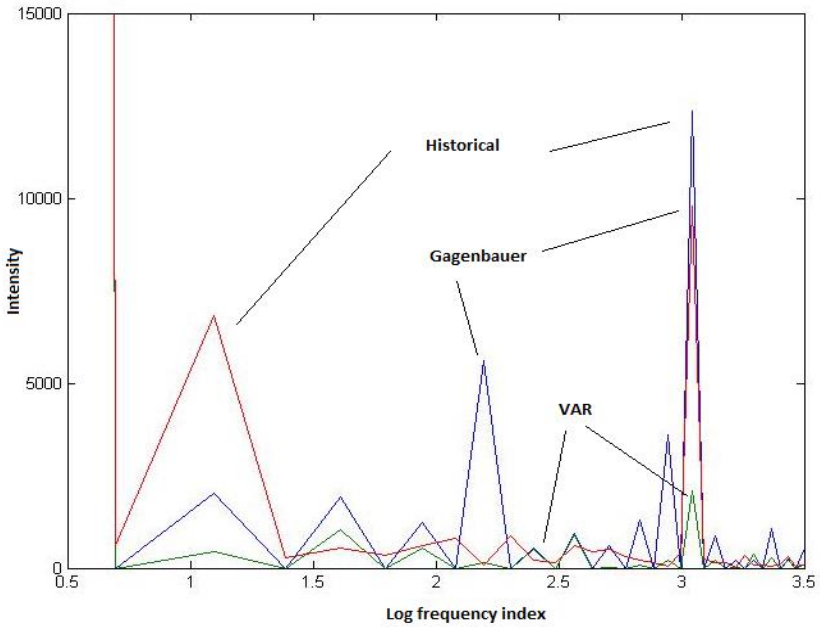

Figure 8.15. Low frequency periodogram segment, log frequency scale, zone 8.

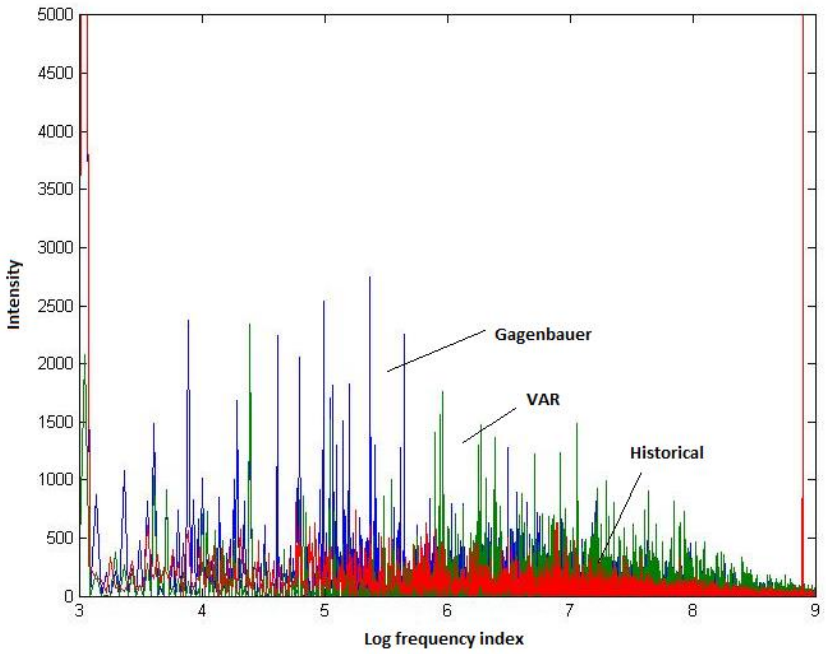

Figure 8.16. Periodogram segment, between annual and diurnal frequencies, log scale, zone 8.

Figure 8.16 shows the periodogram segment between the annual and diurnal peaks for zone 8. It shows that both synthetic series have too much spectral intensity in this range, the Gagenbauer model being worse. This was the case for all sampled zones (1, 8, 17, 20), but was

most pronounced for zone 8. At the far right of figure 8.16 is the diurnal peak – going well beyond the vertical axis' scale. It was found that for each zone, all three peak heights are in excellent agreement, well as the several harmonic peaks found in some zones.

8.4.3 Inter-Annual Variability and Long-Term Means

This final section of analysis on wind speeds looks at various aspects of inter-annual variability in the series, again examining the extent to which the Gagenbauer series outperforms the VAR one. It also looks again at long-term summary statistics.

The first aspect examined is the variance of means, with January, July and annual means calculated. In the case of January means, it was found that variability is generally too large in the Gagenbauer series and too small in the VAR series. January mean variability in the Gagenbauer series is too large for all zones except 1 and 8, with an average of 36% over-estimation and a MAE (i.e. deviation from a ratio of 1) of 0.38. The range of variance ratios, i.e. the synthetic variance/ historical variance, is 0.89 (zone 8) to 1.70 (zone 11). For the VAR model, January variability is under-represented for each zone, with an average of 52% under-representation and a MAE of 0.52. The ratio range is 0.29 (zone 13) to 0.90 (zone 17). It can therefore be said that the Gagenbauer series is the most realistic in this respect.

Considering July means, the Gegenbauer series again generally has too much variability while the VAR is a mixture, in this case outperforming the Gagenbauer series. The Gagenbauer series has excessive variance for all zones except 17, with a mean over-representation of 66% and a MAE of 0.68. The ratio range is 0.96 (zone 17) to 2.48 (zone 4). For the VAR series, variability is under-represented in 11 zones but overall there is a 9% over-representation and a MAE of 0.24. The ratio range is 0.81 (zone 11) to 2.01 (zone 5). It is not entirely surprising that the Gagenbauer series is less realistic for July, since it was observed in chapters 2 and 6 that inter-annual variability is significantly greater in winter than summer, but that incorporating this fact into the model might be too challenging.

The variability of annual means is under-represented in both synthetic series, but more so in the VAR series. In the Gagenbauer series, variability is under-represented in 11 zones, with an average under-estimation of only 2% but a MAE of 0.23. The ratio range is 0.56 (zone 1) to 1.38 (z12). In the VAR series, variability is under-represented for all zones except 16, with an average under-representation of 37% and a MAE of 0.34. The ratio range is 0.47 (zone 1) to 1.10 (zone 16).

Long-term annual means are slightly too large in both synthetic series – by an average of 4% in the Gagenbauer series and 3% in the VAR series. The zonal annual means ratio range for the Gagenbauer series is 1.02 (zone 3) to 1.13 (zone 10), while for the VAR series it is 1.01 (zone 15) to 1.06 (zone 10). Long-term means for January are all too large in the Gagenbauer series – by an average of 8 %, with ratios ranging from 1.03 (zone 13) to 1.2 (zone 3). The situation is slightly worse for July, with an average over-estimation of 14% and ratios ranging from 1.05 (zone 1) to 1.35 (zone 10). The VAR series performs well in this respect, slightly too small for January and too large for July. The January mean is a 3% under-estimate and the range is 0.9 (zone 10) to 1.04 (zone 19). The July mean is a 7% too large, with the ratio range 0.99 (zone 13) to 1.19 (zone 10).

Long-term variances are mostly too large in both synthetic series, with VAR most accurate for January and Ganenbauer best for July and annual mean variances. January variances are over-represented for each zone except 20, with an average 11% over-estimation and a ratio range of 0.92 (zone 20) to 1.35 (zone 2). For July, variances are too large by an average of 27%, with the ratio range 1.10 (zone 20) to 1.56 (zone 10). For the annual variance, the mean over-representation is 12%, with the ratio range 1.01 (zone 20) to 1.26 (zone 10). For the VAR series, January variances are too small for 14 zones, under-representing by an average of 6%, with the ratio range 0.83 (zone 20) to 1.11 (zone 19). July variances however are seriously over-represented in the VAR series – by an average of 47%, with the ratio range 1.31 (zone 3) to 1.70 (zone 10). Annual variances are somewhat better, with a mean 21% over-estimation and a ratio range of 1.11 (zone 20) to 1.36 (zone 10).

The variance of monthly variances was also examined, finding that in this respect the VAR series is more accurate. This variance is generally over-represented in the Gagenbauer series, by an average of 63% for January. The synthetic/ historical ratio range is 0.93 (zone 17) to 2.41 (zone 14), with a MAE of 0.64. For July the over-representation is 73%, with a ratio range of 0.95 (zone 1) to 2.81 (zone 11) and a MAE of 0.76. For annual variances there is an average over-representation of 24%, with a range of 0.63 (zone 1) to 2.73 (zone 11) and a MAE of 0.44. For the VAR series, the variance of January variances is under-represented for most zones, by an average of 15%, with a ratio range of 0.38 (zone 20) to 1.55 (zone 2) and a MAE of 0.26. For July there is a significant over-representation, by a mean of 71% and with a ratio range of 0.73 (zone 17) to 3.34 (zone 9) and MAE of 0.77. The majority of annual variance of variances are under-represented, with an average difference of 6% and a ratio range of 0.48 (zone 20) to 1.69 (zone 9).

In chapter 6 is was noted that positive linear correlations exists between monthly averages and monthly standard deviations, when considering power transformed and de-seasonalised series. It was found that for actual wind speeds these correlation are stronger, with correlation coefficients ranging from 0.71 (zone 19) to 0.83 (zone 1). It was found that both synthetic series recreate this aspect well, with the Gagenbaur series best. For that series, correlation coefficients are slightly too large for the majority of zones, by an average of 5% and with a ratio range of 0.91 (zone 20) to 1.17 (zone 3). For the VAR series, the coefficients are two small for all but 3 zones, by an average of 13% and with a ratio range of 0.55 (zone 20) to 1.06 (zone 19).

Finally, it was observed in chapter 6 that individual months display a wide range of skewness values, and the extent to which the synthetic series represent this was examined. This was done by calculating the range of skewness values for each zone in each series, then calculating the synthetic / historic ratios for each zone. The mean ratio value for the Ganenbauer series is 1.03 and the range across zones is 0.55 (zone 18) to 1.79 (zone 17). For the VAR series the average ratio is 0.94, with zones ranging from 0.15 to 1.27. The Ganenbauer series is therefore the best, on average, with regard to this aspect.

8.5 Conversion to Power Outputs for an Example Scenario

For the final section of the simulation and analysis aspects of this project, the Gagenbauer series were converted into power outputs for an example capacity distribution scenario. The true complexity of converting single-location wind speeds to regional power outputs, and the very approximate nature of any attempt to do so, has been discussed at length in this report. However it was felt that the omission of power output data would be highly conspicuous in an electrical engineering research project. This section outlines how the speeds were converted and the example capacity distribution scenario chosen.

8.5.1 Converting to Hub-Height Wind Speeds

The first aspect to be addressed in converting from 10m Met Office station wind speeds to more realistic ones for wind turbine heights and locations is diurnal seasonality. As discussed in chapters 2 and 4, diurnal seasonalities are not present at offshore locations and reduce significantly with height at onshore locations, with estimation of the extent of decrease well beyond the scope of this project. As described in Chapter 4, Sturt and Strbac in [148] tackle this problem by simply reducing the deterministic seasonality present by 75%, based on a rough guess, and a similar approach is adopted here. During the final stages of series simulation the diurnal seasonality is normally added, before reverse power transformation – the idea tried for the current application was to add only half of the seasonality. Since raising the series to the power 2.5 follows, the overall effect will be to reduce the seasonality by a factor greater than a half. Initially both the standard deviation and mean patterns were halved, and the effect on marginal distribution summary statistics examined for a selection of zones.

A universal pattern was found: the (already small) mismatches between historic and simulated means nearly vanish, skewness and kurtosis values are much closer to historical, but the standard deviations were roughly halved. Clearly, multiplying to correct the standard deviation would lead to means that are too big by a factor of two, so these series are not acceptable. An alternative is to keep the seasonality in standard deviation and halve it for the mean. In this case, the quality of fit for the mean and standard deviation are almost identical to the fully seasonal series, while the skewness and kurtosis values are slightly worse. The problem in relation to the latter two statistics is that the distributions have become more heavy tailed, but this effect would have relatively little effect on associated power outputs, so that this approach to diurnal seasonality reduction seems acceptable.

The next necessary step is to account for the superior resource at realistic wind farm locations within the zones. This is done here through linear speed-up ratios, specific to each of four location types within each zone, purchased by Bath University from Garrad Hassan in order to produce a report [27], as described in Chapter 4. The ratios range from 0.93 for coastal wind farms in one zone to 1.81 for offshore farms in another, but are not published here due to potential commercial sensitivity. For the 9 zones where the choice of Met Office station is different to the report, ratios were adjusted, taking account of the mean wind speed at the previously chosen stations. Speed-up ratios from 10m to a presumed hub-height of 80m were also obtained from the same source, specific to location type and ranging from 1.20 to 1.48. Following the speed increases, values are rounded once again to the nearest knot, to ease the power transformation process.

8.5.2 The Generation Capacity Scenario

The scenario developed here for the distribution of wind generation capacity is based upon the wind farm projects that are currently operational, being constructed, have received planning consent and that have been submitted to the planning authorities. The scenario is for an unspecified time in the fairly near future when all wind-farms in the first three categories have been built, along with a random selection of the capacity in the fourth category. The information was obtained from *Renewable UK*'s wind energy database [2], and values used here were accurate at the beginning of January 2013. The host website allows all projects to be viewed on an interactive map, and this was used to examine each project, determining the zone and location category in which it belongs.

As was previously decided, onshore meteorological station data can only model offshore farms that are close to the coast. The criterion for closeness adopted here is that the wind-farm lies mainly within UK territorial waters, i.e. that is within 12 nautical miles (22km) from the coast. Additionally, projects with total capacity below 25 MW were not analysed, as a means of making this task more manageable.

Table 8.1 below presents the summed capacities for the first three development stage categories, separated according to zone and location type, while table 8.2 does this for the submitted projects. The random allocation of capacities from the latter table was achieved by multiplication of each element with a $U(0,1)$ random variable. The only exception was a 1000 MW offshore project in the Moray Firth named Beatrice, which received backing by the

relevant local authority in June 2013 and therefore seems highly likely to go ahead. The final distribution of capacity in the scenario is presented in table 8.1.

It can be seen that capacity is spread very unevenly between zones, ranging from 2,915 MW in zone 11 to only 2.6MW in zone 18. Nearly 40% of capacity is located in only 3 zones: 3, 8 and 11; with zones 3 and 11 dominated by offshore capacity and zone 8 having the largest onshore capacity. Overall, offshore projects account for just over a third of capacity in this scenario – it would be considerably greater without the distance limitation, with an additional 8,904.0 MW submitted in this category. Additionally, projects excluded due to having a capacity < 25 MW form a total of 1439.3 MW.

	Offshore	Coastal	Lowland	Upland	Total
1	0.9	385.6	0	0	386.5
2	0	43.2	271.2	0	314.4
3	10.0	233.4	420.6	251.6	915.5
4	0	79.0	219.0	356.6	654.6
5	0	3.3	246.6	363.0	612.9
6	7.0	15.1	239.6	539.6	801.3
7	0	90.9	15.0	74.1	179.9
8	0	119.9	921.5	1039.1	2080.4
9	180.0	30.1	391.5	540.2	1141.8
10	62.1	46.5	537.6	61.6	707.8
11	1812.2	133.3	234.0	337.9	2517.3
12	429.0	129.6	480.1	28.6	1067.3
13	464.4	6.0	102.2	0	572.6
14	0	65.5	165.7	435.6	666.8
15	0	0	201.0	0	201.0
16	892.8	41.4	404.0	0	1338.1
17	0	4.5	223.6	0	228.1
18	0	0	2.6	0	2.6
19	0	9.2	6.3	0	15.5
20	1562.8	77.8	4.3	0	1644.9
Total	5,421.2	1,514.2	5,086.3	4,027.7	**16,049**

Table 8.1. Total capacity at each location type, in each zone, that was operational, under construction or consented on 06/01/2013.

	Offshore	Coastal	Lowland	Upland	Total
1	0	0	0	0	0
2	0	0	0	0	0
3	1000.0	51.0	352.9	399.9	1803.8
4	0	0	89.5	177.0	266.5
5	0	0	54.6	525.1	579.7
6	0	0	0	103.0	103.0
7	0	0	0	79.0	79.0
8	450.0	74.7	512.6	619.0	1656.3
9	0	0	560.2	81.8	642.0
10	99.9	39.0	0	0	138.9
11	0	0	114.2	638.7	752.9
12	0	27.0	132.5	0	159.5
13	0	0	54.0	0	54.0
14	0	0	36.0	303.3	339.3
15	0	0	34.5	0	34.5
16	0	33.0	0	0	33.0
17	0	0	0	0	0
18	0	0	0	0	0
19	0	0	0	0	0
20	51.0	0	0	0	51.0
Total	10,505	225	1,941	2,927	**15,597**

Table 8.2. Total capacity at each location type, in each zone, under planning consent consideration on 06/01/2013.

Zone	Offshore	Coastal	Lowland	Upland	Total
1	0.9	385.6	0	0	386.5
2	0	43.2	271.2	0	314.4
3	1010	258.6	719	367.9	2355.5
4	0	79.0	292.2	393.4	764.6
5	0	3.3	272.3	616.8	892.4
6	7.0	15.1	239.6	637.4	899.1
7	0	90.9	15.0	97.6	203.5
8	124.1	173.7	1034.7	1230.5	2563.0
9	180.0	30.1	557.5	612.1	1559.7
10	66.1	83	537.6	61.6	748.3
11	1812.2	133.3	251.4	718.5	2915.4
12	429.0	153.8	535.6	28.6	713.3
13	464.4	6.0	148.7	0	619.1
14	0	65.5	171.7	679	916.2
15	0	0	226.5	0	226.5
16	892.8	44.3	404.0	0	1341.1
17	0	4.5	223.6	0	228.1
18	0	0	2.6	0	2.6
19	0	9.2	6.3	0	15.5
20	1602.5	77.8	4.3	0	1684.6
Total	6589.0	1659.9	5913.8	5443.4	**19606.1**

Table 8.3. Allocation of capacity by zone and location type in the example scenario.

8.5.3 Analysis of the Power Output Series

The linearly increased wind speeds were transformed into power outputs using the same 'softened' power curve that was described in 8.2.2, which accounts for subtle differences in wind speeds experienced across a single wind farm. To guide further smoothing, accounting for thenfact that there are several wind farms within each zone, the aggregated series was analysed in relation to the empirical observations made by Holttinen in [55] – as described in chapter 4, section 4.5. Another publication by Holttinen described in that section is [147], which provides an algorithm for making more realistic the up-scaling of single location wind speeds to regional power outputs resulting from several wind farms. In this section, series generated from a modified version of this simple algorithm are also included in the analysis.

One aspect of the algorithm described is generating the softened power curve. For this project, such a curve was not generated from data, since it relies upon values such as turbulence intensity that are not known for the sites concerned, rather the example shown in the paper was examined and roughly reproduced. The other aspect of the algorithm was to replace the hourly wind speeds with moving block averages calculated from the series – centred around the time in question and of a length given by the dimension of the area divided by the long term average wind speed. In the case of the current GB zones, this would typically lead to averages taken over 7 or more hours. This seems to be too long, since adjacent zones show a peak in cross-correlations at typically 1 or two hours lag, indicating that wind speed signals take this length of time to cross zones, rather than 7 hours. Since we have a multivariate model, it might be the case that instead of moving averages from within the same zones, the average should use up-wind neighbours for past values and down-wind neighbours for future values. However, the progression of wind speed changes across GB may happen in any direction, the prevailing wind direction probably not being dominant enough to justify such an approach.

For these reasons, blocs of 3 hours from within the same zone are considered most suitable here for the majority of zones, whilst the small zones should remain unchanged – i.e. zones 1, 2, 7, 19 and 20. In the paper, block average speeds were calculated and then converted to powers; this seems inappropriate due to the highly nonlinear nature of the power curve. The method here, therefore, is to first convert to powers before averaging. The decision was made not to account for turbine availability here, since future values, particularly offshore, are not well known. Zonal outputs were then aggregated and expressed as percentage load factors. In the paragraphs that follow, the series with moving average applied will be referred to as series 2, the original is series 1.

The mean load factor for both versions of the series is 37.66%, and the median 35.57% and 35.55% for series 1 and 2 respectively. These are slightly higher than average observed values for GB [4], but given the 100% availability assumed here, and the higher than historical proportion of offshore capacity, that is to be expected and the value seems accurate. The series' standard deviations are 21.90% and 21.64%, respectively, making their s.d./mean ratios 0.58. According to Holttinen, this is the value displayed by a country the size of Denmark, while for GB the value should probably be closer to 0.4. This could be evidence that more spatial smoothing should somehow be introduced, although it may also be due to the fact that most wind capacity in the scenario is concentrated into a small sub-set of zones.

Figure 8.17 shows a segment of the modelled aggregated power, for the first 5 days of January, with both versions included. It can be seen that the differences between the two are quite subtle. For a broader view, figure 8.18 shows a year-long segment, with only series 2 included for clarity. On this time scale, the aggregated power output seems to experience very rapid changes and periods of high volatility.

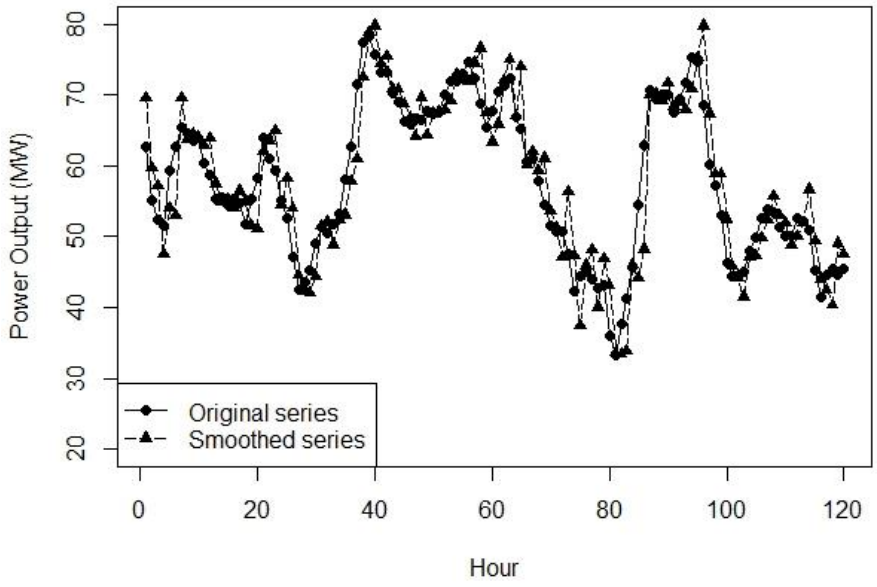

Figure 8.17. A segment of the modelled aggregated wind power output – with and without additional smoothing. 1st 5 days of January.

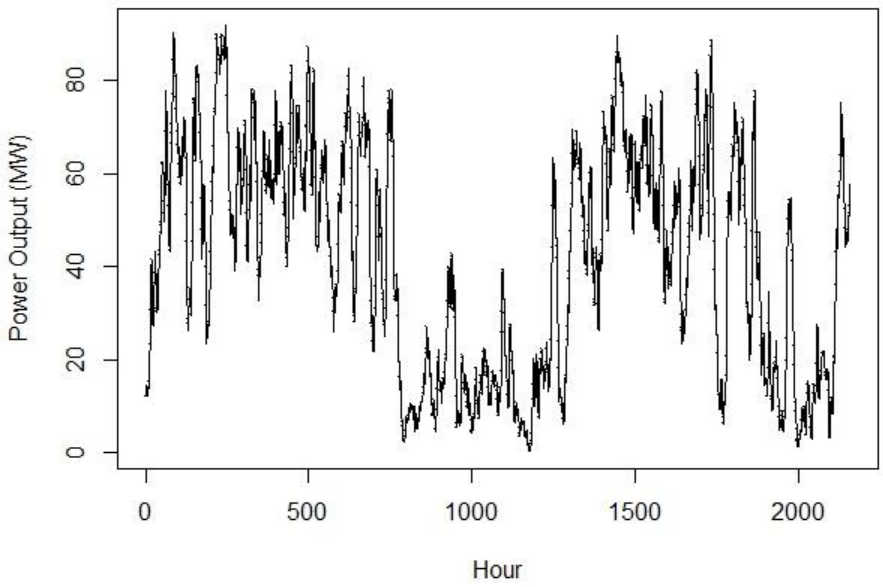

Figure 8.18. A year-long segment of the modelled aggregated wind power output, with additional smoothing.

Holttinen states that load factors of 100% are not realistic, rather the maximum should be about 85% to 95%, depending on the size of the country. For series 1, the maximum observed load factor is 97.25%, with 95% exceeded during 12 hours only (0.01%). Results are very similar for series 2, and seem entirely acceptable, particularly given the assumed 100% availability. With regard to minimum output, Holttinen states that load factors less than 1% should be non-existent for a large country, while occurring less than 5% of the time for a small country such as Denmark. For both series, the minimum load factor is 0.09, with factors below 1% happening 0.34% and 0.28% of the time, respectively, which are good results.

It is stated by Holttinen that hourly load factor changes should be < 20%, with a standard deviation < 3%. This is not the case for series 1, with a maximum change of 28%, although changes exceeding 20% only occur between 0.07% of hours. The standard deviation is significantly too large at 5.17%, but again could be the result of the concentration of capacity in a few zones, rather than a fault of the model. Series 2 does better however, with a standard deviation of 3.18% and a maximum change of 16.15%.

Overall, both series have performed well, particularly series 2 with regard to hourly changes.

8.6 Chapter Summary

The first part of this chapter outlined the algorithms developed for simulating the rather complex model developed in chapter 7, including problems that were faced and methods developed, mainly through trial and error, to deal with them. Some problems were known from the outset – such as reproducing a high dimensional joint distribution, where the marginal distributions deviate significantly from normality. Other problems were unexpected, such as exploding variance.

Throughout the simulation process, many censoring and re-scaling steps unfortunately had to be included so that the synthetic series did not deviate unrealistically far from the historical series – although admittedly there was a subjective element to judging what unrealistic should be. Despite such restraints, the final series for many zones ended-up with considerably higher values of skewness and kurtosis than their historical counterparts. That flaw aside, comparisons between individual series were mainly very favourable, and the model appears to be successful.

The joint distribution of synthetic wind speeds was also shown to be a good match to the historical series, although some deviations from joint-normality in the latter may be seen that are naturally not reproduced in the synthetic series. An approach was therefore developed that incorporated a novel model involving a transition matrix between vector states of both wind speeds and volatility. This approach made very modest improvements to the synthetic joint distribution.

Further detailed analysis of individual series was then presented, with a simpler model also developed and analysed that ignores the long memory and heteroskedasticity aspects of the original. From a comprehensive series of tests, it may be concluded that both models have their strengths and weaknesses, the more complicated model being more accurate most often. The tests certainly illustrated that the model is not perfect, but the impact of those imperfections on its ability to deliver accurate values for e.g. capacity credits is unknown.

A scenario was developed to produce power outputs that are realistic in many respects.

Chapter 9

Conclusions and Further Work

This chapter presents the conclusions of the research. It begins with a full review of the research objectives achieved, referring to the precise objectives set out in Chapter 1. After that, possibilities for future work are proposed and finally a summary of conclusions is presented.

9.1 A Review of Objectives Achieved

This section provides a detailed breakdown of the research objectives achieved, mirroring precisely the objectives presented in Chapter 1.

Chapter 2:

A literature review was conducted, examining both the wind speed field (with an emphasis on GB) and how local wind speeds relate to a much bigger picture in terms of atmospheric circulation.

2.1: A thorough review of wind resource meteorology was presented.

2.1.1: This section explained how the excellent wind resource in GB derives from the frequent passage of low pressure systems. However wind speeds are often determined by the presence of several low and high pressure systems within a larger area.

2.1.2: An hourly temporal resolution for wind speeds was shown to be suitable – fine enough to reveal the details of synoptic scale variability, but long enough so that turbulence is smoothed out.

2.1.3: It was seem from the literature that wind speeds are typically very well represented by Weibull distributions, and sometimes a special case – the Rayleigh distribution. At other times the distribution of wind speeds may be bi-modal and a Gaussian mixture distribution is more appropriate.

2.1.4: Simple approaches were presented for increasing wind speeds to account for the greater height of wind turbines, compared to observation stations. It was however found that

such approaches are only accurate at modelling the long term average relationship. It was concluded that this research project should not aim to include a realistic model for wind shear.

2.2: A literature review was provided of the spatiotemporal properties of the wind resource – both in terms of wind speeds and wind farm outputs, including detailed studies on the GB resource.

2.2.1: The work of Dr Graham Sinden on the GB wind resource was presented and discussed, based upon the recorded wind speeds at between 40 – 60 locations for each modelled hour over a number of years. Many insights were discussed, relating to e.g. the maximum spatial extent of high and low wind speeds, the variability of annual capacity factors and the considerable variability of aggregated wind availability during peak electricity demand hours. It was shown that spatial correlations follow an $e^{-k \cdot d}$ structure (where d is distance), albeit with some considerable deviations from that pattern.

2.2.2: A number of technical reports by the United States Government's National Renewable Energy Laboratory (NREL) were discussed. These have the rare property among the literature that they considered temporal resolutions starting at 1 second. It was concluded from these reports that sub-hourly variations in power output should largely cancel out when aggregating over a large number of wind farms.

2.2.3: A paper was discussed noting that some future scenarios involving very large wind energy capacities are necessarily unrealistic due to atmospheric energy conservation principles.

2.3: A literature review was presented on the meteorological practice of synoptic classification – i.e. the placing of large scale (roughly 1000 km x 1000 km) patterns of atmospheric circulation into useful and meaningful categories.

2.3.1: An introduction to the relevant concepts and crucial literature for synaptic classification was presented.

2.3.2: A detailed description of one synoptic classification system was provided – the Grosswetterlagen (GWL), along with this system's insights about peak electricity demand times. The main insight in there is that the majority of truly peak demand times correspond to circulation situations known as blocking, rather than a high pressure system over GB. A very useful circulation type catalogue was presented, the HB-GWL, along with re-analysis datasets, which could potentially be used as a training dataset for the model, rather than weather station observations, and a useful variant on the GWL system involving a much reduced number of types (referred to as reduced GWL).

2.3.3: A literature review was presented on the objective classification of surface wind measurements – seen as a potentially useful tool, although perhaps only so for the modelling of wind vectors (i.e. speed and direction).

2.4: A meteorological literature review was presented on the nature of climatological variability. Ensuring that very low frequency variability is accurately captured and reproduced was identified as a priority for the model and simulation methodology to be developed.

2.4.1: Descriptions found in the literature were presented of inter-annual and inter-decadal variability in mean wind speed and mean wind power load factor.

2.4.2: A review of the meteorological literature on long-term spatiotemporal patterns in atmospheric circulation in the UK and Europe was presented. This included the Lamb circulation classification system, principal component analysis (PCA) and the North Atlantic Oscillation Index (NAOI). The main insight was that there occur clusters of years where certain types (particularly Westerlies) are more common than their long term average, and their frequency of occurrence is correlated to the NAOI. Within GB, some locations are much more strongly affected than others by such long-term variability.

2.5: A review was prepared of alternative mathematical approaches for representing climatological non-stationarity.

2.5.1: A discussion was provided of the concept of long memory in time series, and its presence in meteorological phenomena - manifesting, for example as the Hurst effect observed on the Nile. Such an approach was presented as desirable in that it avoids the need for somewhat artificially imposed transitions between a discrete number of states, with artificial minimum lengths. Two examples were found in the literature - one modelling wind speeds as having a deterministic annual seasonality and long memory, the other as having seasonal long memory (with an annual period).

2.5.2: A brief discussion was provided of conditional heteroskedasticity and its presence in wind speed time series.

Chapter 3:

Chapter 3 consisted of a presentation of the fundamental mathematics of time series analysis and modelling, providing a reference for future chapters. This included the fundamental tools of analysis, model structures commonly used and popular model fitting techniques.

3.1: The basic concepts of ARMA Models were presented.

3.1.1: The basic concepts, notation and terminology associated with time series analysis and ARMA models were presented. Correlograms and partial-correlograms received considerable attention.

3.1.2: The representation of time series in the frequency domain was presented, with topics ranging from periodograms to the transfer functions of ARMA models.

3.1.3: The relevant theory of SARMA seasonal models was presented, including their ACF and spectral densities – the latter predicted as being very 'messy' in appearance for a model with annual seasonality.

3.1.4: Concepts associated with the differencing of time series were presented, including how to judge that it is necessary. The very general ARIMA and SARIMA models structures were presented.

3.1.5: Methods for transforming the probability distribution of time series were presented, including Box-Cox transformations and methods involving inverse cumulative distribution functions.

3.2: ARMA model extensions were presented that allow representation of long memory and heteroskedasticity.

3.2.1: ARMA model extensions were provided that allow for long memory – both 'regular' and seasonal, i.e. ARFIMA and GARMA models, respectively. Both models involve fractional differencing, and GARMA models define series known as Gagenbauer processes. The fractional differencing operators associated with each model were presented as infinite polynomials (that may be truncated) of the back-shift operator. The operators have different forms depending on whether one wishes to fractionally difference a series of observations, or simulate a long memory series with a random number generator. A recursive formula was presented for the definition of polynomial coefficients for GARMA models.

3.2.2: Univariate models for conditional heteroskedasticity were presented: ARCH, GARCH and further generalisations – including APARCH and ARCH-in-mean. Multivariate extensions were also presented. It was seen that such models are very powerful, capable of capturing many complex dynamic behaviours, but that they are extremely challenging to fit, particularly for a large number of dimensions.

3.3: The basic theory and practice of order selection and parameter fitting for ARMA type models was presented.

3.3.1: A conceptual introduction to the fitting of ARMA models was provided, with an emphasis on correlograms and partial-correlograms.

3.3.2: The Yule-Walker equations and the Levinson-Durbin algorithm were presented as essential tools. Also, a multivariate extension by Whittle was presented – demonstrating that such an extension is not as straightforward as one might initially imagine.

3.3.3: The crucial role of the BIC and AIC information criteria in model order selection was discussed, including their role within the Hannan-Rissanen procedure – which proved central to the model fitting process adopted in this research.

3.3.4: The theory of maximum likelihood estimation (MLE) was presented, including in the multivariate case. The difficulty and computational expense of MLE in the multivariate case was immediately obvious.

3.3.5: A number of tests were presented that may be applied to model residuals to check if the model structure is adequate.

3.4: A brief introduction to the theory of wavelet transforms was provided, along with a discussion about their potential use in modelling the wind resource. Examples from the literature were provided of their use in estimating and simulating long memory processes, along with hierarchical models involving hidden Markov chains. In the multivariate case, it was seen that wavelet analysis is very powerful, with the ability to resolve differing cross-covariances on different timescales. However, with that power comes considerable complexity – and it was judged that this complexity might be unmanageable for a 20-dimensional problem.

3.5: A discussion was provided on non-Gaussian multivariate processes, and the use of copula functions as a means of going beyond the multivariate normal assumption. It was once again seen, however, that such modelling may be impractical for a 20-dimensional problem.

3.6: In introduction to the theory of Markov Chain Monte Carlo models was provided, with a discussion of their potential benefits – mainly that the direction and scale of wind speed changes may be conditional on the 'current' wind speed value. The presentation included extensions to the basic model type, including higher orders and multivariate models. One major shortcoming of multivariate extensions is that while they allow the 'future' state at a given location to depend on the 'present' state at a number of locations, they cannot allow for realistic spatial correlations in the changes/ innovations between the 'present' and 'future'.

Chapter 4:

Chapter 4 was a second literature review, reporting on efforts to date to create and apply time series models to wind speeds. This included previous work at the University of Bath that acts as a starting point for the model developed here. Drawing upon chapters 2 and 3, the most advanced modelling work conducted to date was discussed and critically evaluated.

4.1: The Bath wind model, which emerged from previous wind speed modelling work conducted at the University of Bath, was presented and critically evaluated.

4.1.1: Present and critically evaluate The Bath wind model was presented and discussed in much greater detail than previously in Chapter 1. It was reported that the model's methodology did not involve transformation of the distribution of wind speeds, nor the removal of diurnal seasonality. These are viewed as partially justifiable given that the model is for winter peak times only, however it means that the outputs cannot be time-stamped in any meaningful way. The presentation included the process of converting the wind speeds initially generated to zonal power outputs, involving speed-up ratios specific to each zone and location type. Speed-up ratios from 10m to hub height were specific to each location type. The availability of each turbine was simulated as a Bernoulli random variable.

4.1.2: The way in which the Bath wind model's performance and suitability were previously assessed was discussed. It was found that most validations were trivial in that they were bound to succeed if parameters were fitted correctly.

4.1.3: Critically review Other wind modelling work conducted at the University of Bath was reviewed, but found not to be slightly inconsistent and not of great relevance to the present research. One important point made in the literature is that the nature of power transformations is inevitably very compromised in a multivariate setting.

4.2: A review was presented of ARMA/VARMA models for wind speed developed by other authors.

4.2.1: A review was presented of previous work that constructed univariate ARMA models for wind speed. Considerable disagreement was found to exist with regard to the optimal model order. A consistent message in the literature is the importance of fitting models to several years of data.

4.2.2: Six examples of previous work that constructs VARMA or similar multivariate models for wind speed were reviewed, with two being particularly noteworthy. One is the VARMA model of GB wind speeds developed by Hill et al. in Strathclyde. This is very similar to Bath wind model, but is an all-year model that assumes complex deterministic seasonalities with diurnal and annual periods. It is worth noting that the modellers: (i) considered OLS parameter fitting

to be entirely acceptable; (ii) found AR(2) to be the optimal order (perhaps as the concern in short term forecasting, rather than simulation); (iii) could only find good quality and long-period observation data at 14 locations; (iv) believe that after removal of seasonality, no power transformation is necessary; and (v) avoid any method of conversion to wind powers, on the basis that any reasonably simple method of doing so will necessarily be highly inaccurate.

The second model is concerned with finding optimal spatial arrangements of wind capacity across Germany, so that the lowest percentiles of their accumulated outputs are as large as possible. To do so effectively, they cannot rely on the assumption of linear dependencies, i.e. Gaussian copulas. Their modelling task was somewhat different, since most of the data available to them was in the form of daily average wind speeds. The modellers remove deterministic seasonalities in both mean and variance, fit univariate AR models and hope that merely spatial relationships remain in the residuals – found to be true for daily averaged data only. Then, they use the method of pair-copula vines to capture the spatial relationships, with different copula families needed for different pairs. This shows that allowing a non-Gaussian copula specification is certainly possible, even in a 40-dimesnional case, but remains extremely challenging.

4.3: Previous work modelling wind speed as a Markov Chain, or a semi-Markov process, was reviewed. Some involved assuming a higher order structure, but none were multivariate.

It was found that the value of increasing the model order, and increasing the number of states, depends on how success is measured. It emerged that regardless of model structure, Markov chain models under-represent persistence beyond a short horizon – despite being able to e.g. effectively predict the length and frequency of extended calm periods. An interesting metric was invented in one article – the sum of differences between historical and simulated ACF values up to a fixed lag (12 hours), and a similar approach was adopted in the current research (within chapter 7). One semi-Markov model of wind speed was reviewed, with performance found to be quite good.

4.4: A review was presented of modelling work that employed hierarchical models of wind speed.

4.4.1: Previous work modelling wind through regime-switching regressive models of wind speed were reviewed. A very strong case for regime switching was made for 10-minute resolution series (with changes mainly effecting variance and persistence), with the case assumed to be less strong at an hourly resolution. The optimum number of states emerged as 2 to 3, corresponding to half the number of reduced GWL states.

4.4.2: The use of wavelet transforms in wind speed field modelling was reviewed, coming to the conclusion that the method is very powerful, but also too complicated for a 20-dimensional model.

4.5: An insightful literature review was presented on the process of producing regional wind power outputs from single-location wind speed time series. A particularly useful article presents key features of the distributions of spatially aggregated power outputs, and how they change with the geographical extent of the aggregating area.

4.6: A literature review was presented on the direct modelling of aggregated wind power outputs. One reviewed article compared a model to available data in several countries, while another produced synthetic data for a hypothetical GB wind fleet (and as such only wind speeds could be used for validation). Synthetic series were subjected to rigorous validation, and performed well – the only shortcoming being that inter-annual variations in distributions are definitely under-represented, although to a rather uncertain extent. An interesting approach is taken to diurnal seasonality in expectation mean at hub-height: eliminate it entirely for offshore capacity and halve it for onshore capacity.

4.7: A review was presented of previous work that models wind speed time series as possessing long memory and conditional heteroskedasticity – with one example found that incorporate both.

4.7.1: Previous work where wind speeds were modelled as either an ARFIMA or a GARMA process were reviewed. In one reviewed article, the first to attempt the representation of wind speed as a long memory process (with limited computational power), a fairly high value was found for the long memory parameter, yet surprisingly low values for short memory persistence. This seems to suggest that one compensates for the other, and that a wide variety of parameter value combinations can be sensible.

4.7.2: A review was presented of work where wind speeds are modelled as having conditional heteroskedasticity. Only one article was of note, which also incorporated long memory by assuming an ARFIMA-FIGARCH model structure. The article's authors note that they could have followed the approach of assuming a stochastic nature for the periodic components, allowing specifications such as GARMA, SARMA or SARIMA models.

The chosen specification allows for long memory in the conditional heteroskedasticity, however it was found that if such long memory is present, it is very weak. Thus it was decided that the possibility of modelling long memory in variance should not be considered within the current research. The article found that the long memory in expectation value was also week, but very likely to be present.

Chapter 5:

Chapter 5 described the process of acquiring the data, processing it and assessing its quality – with very favourable results. The end product of the data cleaning and filling processes was a 20 year series for 20 locations across GB, with a wind speed value present for each hour. This dataset cannot be freely distributed, due to much of the data being identical to that obtained under restriction from the BADC. However, analysis of the series may be openly shared.

5.1: Example wind speed observation datasets were obtained via the BADC, and converted to a format that's easy to manipulate in *Matlab* (the software choice for almost all the calculations involved in this research). Although deemed suitable overall (and preferable to a reanalysis dataset), assessment of their quality revealed many problems, including missing hours, several values for one time stamp and general temporal disorder.

5.2: The example wind speed datasets were cleaned – establishing what should be done came first, followed by the development of algorithms to do it.

5.2.1: An effective clean-up algorithm was developed that ensured correct chronological ordering, identified gaps and made a choice between multiple entries for the same hour, when they occurred. The algorithm was largely successful, although it occasionally left a few large blocks in the wrong location within the series, thus making it necessary to check all series by visual inspection.

5.2.2: Wind speed distributions for the example series were explored. At some locations, they were found to be almost perfectly Weibull distributed, whilst others were bimodal, with the smallest peak at low wind speeds of 2 – 3 knots (1- 1.5 m/s). Others had a very significant excess of 0 knot wind speed recordings.

5.2.3: Options for the potential removal of erroneous readings of 0 knots were investigated, although no successful method of distinguishing between genuine and erroneous values was found.

5.3: A suitable combination of Met Office station locations were selected, one for each zone, for which a high quality 20 year sample was present. The period was the 1st of January 1988 to the 31 of December 2007, and the datasets for this period were obtained and cleaned.

5.3.1: Many MIDAS observation stations were explored to find the best combination of 20, on the basis of multiple criteria – some relating to the individual stations and others to the entire set. It was established that for a selected 20-year period (1988-2007), stations could be found in each zone with a high percentage of data present after clean-up (>95%), and with no gaps longer than a few weeks. This was not true for longer periods (with 25 years was attempted).

Nine of the chosen stations differed from those of the Bath model, presumably as a result of a more rigorous selection process adopted here.

5.3.2: The quality of data was assessed in detail for the final selection of zones, following clean-ups, including the distribution of gap lengths. The results indicated that the datasets are of a high quality.

5.3.3: *Google Maps* and *Google Earth* were used to characterise the precise locations of the Met Office station masts, with no serious reasons to doubt the data quality found.

5.4: All gaps in the series were filled in, as some relevant *Matlab* functions require this. In order to achieve this, the series were transformed so that they roughly form a multivariate normal, with standard normal marginal distributions.

5.4.1: The series were Box-Cox transformed – after exploration and consideration of the optimal way of doing so. Subtraction and division then followed, to render the set of series approximately a multivariate normal with N(0,1) marginals. Since different power transformations are suitable for different series – it was necessary to identify the important ones and fond the best compromise among them. The bimodal nature of several distributions was also problematic, causing negative skewness coefficients.

5.4.2: A successful algorithm was developed to remove deterministic diurnal seasonality in mean and variance. This involved fitting a non-parametric daily pattern for both, with that pattern varying smoothly from day to day.

5.4.3: A method for filling-in missing was developed, based on the exploration of several interpolation and forecasting approaches. One approach tested was highly sophisticated, involving univariate ARMA and a VAR model and even forecasting with time reversed. The other approach simply involved spatial interpolation and was found to work best for all gap lengths.

Chapter 6:

Chapter 6 presents a detailed statistical analysis of the historical series, confirming and enhancing some of the insights from chapter 2. It contains an exploration the usefulness of transformation of the series into principal components and also how the components relate to the reduced GWL atmospheric circulation classification scheme. Based on this analysis, a concept was presented of how wind speeds and electricity demands may be realistically coupled in Monte Carlo simulations.

6.1: A statistical analysis of the wind speed series was conducted, with diverse aspects of wind resource dynamics examined.

6.1.1: Summary statistics for each series were presented and discussed, along with their cross-correlations. Expected patterns were found, such as the north west of GB being the windiest, and that mean wind speeds are positively correlated to long-term variance.

Before Box-Cox transformation, all zones had large positive skewness coefficients. The transformation brought the skewness coefficent of most, but not all, zones closer to zero. This is interpreted as becoming more Gaussian-like. The exceptions were partly due to compromised nature of the power parameter, but also due to the 2nd low-speed peak.

All cross-correlations are positive, with some very large. Summing them for all zones confirmed that no location is particularly isolated.

6.1.2: Time series segments were prepared and discussed, with wind speeds averaged over hourly, monthly and annual resolutions. Examining the plots revealed a degree of self-similarity on the different timescales, consistent with the presence of long memory. The annual seasonality was noted as appearing stochastic in nature, or at least having a strong stochastic component. Examining 2 example years of raw data for zone 1, the monthly variance was noted to vary by a factor of 10, reducing to a factor of 4 after power transformation.

6.1.3: Several figures were produced, exploring differences between individual months, in terms of 3 summary statistics: mean, variance and skewness. These figures were scatter plots that present the joint distributions of 2 monthly summary statistics, and it would be a very strong validation for the synthetic series if they can reproduce such joint distributions, despite the fact that they cannot be built into the model explicitly. It was noted that skewness coefficients display considerably variability, much more than would be expected from a stationary process.

6.1.4: A selection of correlograms and partial-correlograms were prepared and analysed, with an emphasis on exploring very long-range lags. It was noted that the autocorrelation function (ACF) has the form of a very slowly decaying sine wave, the amplitude being >> the white noise boundary. Also, the centre is significantly above zero. All these observations point towards the presence of long memory, and seem consistent with the Gagenbauer process as the most accurate model.

6.1.5: A selection of periodograms were prepared and analysed, some plotted with logarithmic axes, others not. It was noted that for some zones, e.g. Valley, a single dominant pole at the annual frequency suggests very strongly that wind speeds there should be modelled as a Gagenbauer process. However, the situation is not so clear in other zones, where the annual frequency remains the maximum, but not nearly to the same extent. It was noted that those

zones there the annual period spike is strongest seem to be those zones identified in the literature (Palutikof et al. [80]), as being the most heavily influenced by changes to the NAOI. This observation was seen as a further indication that the Gagenbauer process is a suitable model, although the argument is by no means overwhelming. In favour of treating the wind speeds as a regular long memory process, it was noted that for several zones the spectral intensity is a straight line for several of the lowest octaves in the log-log periodogram plot – as is expected for a regular long memory process. A further plot provided strong evidence that the extent of long memory is itself seasonal, but it was decided that capturing that aspect of dynamics is too challenging for the current research.

6.2: Principal components analysis was conducted on the Box-Cox transformed and completed series; with the results presented and discussed. It was noted that the 1st component is essentially a spatial mean and accounts for 43% of the variance. As the order number increases, the component series become more like volatile local oscillations, combining positive and negative weights from a limited selection of zones. If one wished to fit models to the component series, it was noted, then each one would have a very different structure.

6.3: A daily catalogue of reduced GWL circulation types was provided by Dr. David Brayshaw of Reading University, which was matched to wind speeds across the 20 year period, to enable analysis.

6.3.1: The 20 year period was first examined in terms of changes in the relative frequency of occurrence for the reduced GWL types. Plots showed there were clearly clusters of 2 to 3 years where each type is more prevalent, but also lower-frequency oscillations with periods on the scale of decades.

6.3.2: Probability distributions were estimated for each of the principle components, during each of the GWL types (with probability distribution being estimated here by the relative frequency of occurrence during the historical period). The distributions were then compared across the reduced GWL types, with differences found to be quite subtle – but most pronounced for the 1st component, referred to as PC1. The frequency of occurrence of central values of PC1 were found to be very similar for each GWL type, however the relative occurrence of tail PC1 values are much higher (as a fraction) for some GWL types than others. One can therefore revert the situation and calculate the probability mass distribution for the reduced GWL's given the PC1 value. For central PC1 values, the masses are similar, but not so for the highest and lowest observed values of PC1.

6.4: The value of applying k-means clustering to the wind speed fields was investigated, with the research question being whether certain clusters correlate with atmospheric circulation

types – or even indicate a significantly higher/lower probability that a certain weather type is occurring. Daily averaged wind speeds were used. The cluster mean values were laid out in spatial relationships roughly representative of the zone's position within GB, for ease of interpretation. Although some patterns look meaningful, the decision was taken, as advised by Dr Brayshaw, that was that this was not a fruitful avenue of research - although the introduction of wind directions might change this.

6.5: A possible method for connecting demand and the wind speed field was proposed – but is merely conjecture. The idea is that a wind speed field series may first be generated. A Markov chain model could then be used to generate a chain of reduced GWL types, or perhaps some alternative system of hidden states. After the tentative state is generated, it may be rejected or accepted, with the probability of doing so determined by the probability mass associated with that state, conditional on the PC1 value of the wind speed field. A model could be developed for electricity demand, obviously involving suitable temporal auto-correlation, but also distributions that are conditional on the reduced-GWL/other hidden variable, as well as the day of year and hour of the day.

Chapter 7:

Chapter 7 chronicled the development of the iterative process of choosing the model structure, establishing optimal parameter values, testing the suitability of those choices and considering possible improvements to the model structure. Although there was no clear evidence for choosing a GARMA model structure over ARFIMA, the decision was made in favour of the former at the beginning of the model fitting process. It was later found that univariate APARCH models were suitable for the conditional variance, leading to the choice of VGARMA-APARCH – initially with one Gabenbauer frequency, then 2 for each zone. It seems that the current research is the first to construct any model with this precise structure, and as such it involved a mathematical leap of faith.

7.1: The most suitable method of fitting the initial choice of model structure for the conditional expectation values was established and applied. That initial choice was an annually seasonal VGARMA(3,2) model, in which fractionally differenced wind speeds – differenced as a process with annually seasonal long memory – are described by a VARMA(3,2) model. It was found that maximum likelihood estimation (MLE) was unsuitable.

7.1.1: A methodology for fitting an annually seasonal VGARMA model to the 20 series, making use of the Hannan-Rissanen procedure, was developed and applied, making no assumptions at this stage about the error variance structure. Separate modelling of conditional expectation

values and conditional variance is not ideal, but is the only possibility for such a complex model structure, and for such a high dimensionality.

The developed process identified the optimal values for the differencing parameters, identified ARMA(3,2) as the best model order, and fitted the optimal ARMA parameters. Examination of periodograms for the fractionally differenced series demonstrated that long memory had been successfully removed.

7.1.2: Having fitted the model through a method that was conceptually quite straightforward, it was decided that attempts would be made to further fit the parameter values through MLE. To inform this process, a detailed literature review was conducted on previous attempts to use MLE to fit the parameters of similarly structured models. It was found that closed-form results are often provided in the literature for Gaussian errors, but also occasionally for other distributions, such as student-t.

A very popular substitution for full MLE in the context of long memory models is Whittle's quasi-MLE method, involving the relationship between periodograms for the series sample and the theoretical spectral densities for the proposed comination of model structure and parameter values.

7.1.3: An attempt was made to use quasi-MLE to establish the parameters of the VGARMA(3,2) model, using previous estimates as starting values. It was established that the method was not viable for a multivariate GARMA process, due to both extreme computational expense and inconsistent results. Results for an univariate MLE trial fit did seem superior to the Hannan-Risannen type procedure, when theoretical power sprectra were compared to periodograms. However, inspection of correlograms for the MLE parameter fractionally differenced series indicated that some long memory remained.

A big problem was noted at this stage: regardless of method, the theoretical peaks at the annual frequency are an order of magnitude too small. This was interpreted as clear evidence that the proposed model structure was inappropriate.

7.2: A discussion was presented detailing how a replacement model structure was proposed. This new structure was a 2-factor-VGARMA model with deterministic annual seasonality. Develop and apply A methodology for deciding upon the order, and fitting the parameters of this model structure was applied and presented.

7.2.1: The complete inability of the theoretical spectra to match the spectral peak in the was interpreted as indicating that a deterministic annual seasonality should be removed, despite its apparently stochastic nature. Indeed, a deterministic annual seasonality was successfully removed, following some experimentation on the optimal extent of smoothing.

7.2.2: Periodograms, smoothed to varying extents, were prepared for the series after removal of the deterministic annual seasonality. It was noted that following the removal of the deterministic climate, the periodograms remained almost unchanged, with the exception that the single, very large spikes were gone. For some zones, the tallest spike in the new periodograms were at the lowest Fourier frequency, so they might reasonably be modelled as a regular long memory process, with the long memory effect light. This is not the case for all zones, however – for some zones, the highest peak occurs at around the 200^{th} Fourier frequency, for example.

It was decided that rather than looking at individual spikes, very heavily smoothed peroidograms should be more revealing. A basic pattern was found that is almost universal across the zones: a clump of high intensity, clearly containing the greatest peak and centred between 1 and about the 100^{th} Fourier frequencies; along with several smaller clusters. One of these smaller clusters is somewhat more prominent, and was consistently centred between roughly the 60^{th} to 400^{th} Fourier frequencies. It therefore seemed more accurate to assume that there are two Gagenbauer frequencies, and model the transformed wind speeds as a multivariate 2-factor-GARMA process. This was also considered advantageous over a single-factor model since the individual δ values will be smaller and the spectral intensity less 'peaky', i.e. more spread-out, in line with the periodograms.

7.2.3: A methodology for establishing the order and parameter values of a 2-factor-VGARMA model was developed and applied. As part of this methodology, the Gagenbauer frequencies selected by visual inspection of the heavily smoothed periodogram. Initial searches found that, for each zone, solutions whereby the higher frequency differencing parameter has values in the region of 0.03 were close to optimal. Therefore, this was taken as a fixed value, for each zone, simplifying the estimation process for the remaining parameters. The theoretical spectra for initially fitted univariate 2-factor-GARMA models were found to be significantly better fits to the periodograms than the single-frequency models were.

More advanced stages of the fitting procedure revealed that a switch to a VGARMA(3,1) model structure should be made, as it is very slightly superior to VGARMA(3,2) with the differencing changed. Fitted values for the low frequency differencing parameter displayed considerable range, from 0.03 to 0.13. It is not entirely clear that the extent of long memory presence in the 20 series truly varies as much as the differencing parameters suggest, although it is certainly possible. It may rather be a quirk of the fitting procedure, essentially caused by a relationship similar to colinearity between long and short memory persistence. Unsuccessful attempts were once again made to re-fit the parameters using quasi-ML.

7.3: Attentions now turned to the conditional variance structure. A detailed analysis was conducted on the residuals of the 2-factor-VGARMA model, with a focus on spatiotemporal associations. The analysis was discussed in terms of indications of a suitable model structure. The APARCH model type was selected as appropriate and a process was developed and successfully applied for order selection and parameter fitting.

7.3.1: A detailed analysis of the 2-factor-VGARMA model residuals was conducted, particularly their spatiotemporal structure, with the main analysis tool being autocorrelation functions. Considering e.g. zone 9, lag-zero cross-correlations (i.e. purely spatial relationships) were found to have small positive values of up to about 0.1. For lags > 0 however, there appear to be few, if any, significant auto- or cross-correlations, with only a few very small negative cross-correlations noticeable. This is seen as a very positive indication that the VGARMA model is a good fit. Examining correlations for squared residuals, it was found that auto-correlations are much bigger than cross-correlations,even though many cross-correlations are significant. Given the undesirability of a multivariate model of conditional variance, due to the very large number of parameters involved, this is taken as sufficient justification for adopting univariate GARCH-type models.

An annual seasonality in smoothed, daily averaged variance was investigated and found to of considerable relative magnitude. It was also found that the inter-annual variability in daily averaged variance is considerably greater for some days of the year than others, and that this feature should be reproduced in the synthetic series.

Examining shorter segments of the residuals series, and squared series, volatility clustering was very obvious. A suggestion of asymmetry was noted, namely that large negative values seem to have a greater effect on the subsequent variance than positive ones.

7.3.2: An algorithm was developed and applied to remove the observed itra-annual and inter-annual seasonality in variance, successfully leaving flat profiles. The task was in fact very challenging, with much experimentation necessary and a complex final algorithm. It was assumed that removal of deterministic seasonalities in both the conditional mean and noise aspects of the wind speed process adequately accounts for the fairly weak positive correlation between mean and variance observed in Chapter 5. In other words, it was assumed without investigation that there is no need for an ARCH-in-mean element of the model structure.

7.3.3: A discussion lead to the conclusion that univariate APARCH(1,1) models of conditional heteroskedasticity are suitable for adequately capturing the remaining volatility clustering behaviour. It was further concluded that a sub-set of such models is appropriate, to ease parameter estimation, i.e. it was deemed necessary to keep the power parameter δ at a fixed

value of 2. A methodology for estimating the model parameters was subsequently developed and applied. It was then demonstrated that the fitted models were successful in removing spatiotemporal associations in variance.

It was hoped that model parameter estimates could be established through maximum likelihood optimization, since the problem is univariate. The empirical distributions of the conditional residuals were estimated through jernel density estimation, and were found to deviate considerably from being Gaussian, on account of excessive kurtosis. Suitable options for which analytical MLE solutions exist were g.e.d. and student-t, with investigations establishing that g.e.d. was the best choice for 15 zones, with the quality of fit varying significantly.

A technical difficulty for using MLE was that the best-fitting distributional parameter values for the original, conditional errors were the only initial estimates available for the unconditional errors' distributions. Initial estimates for the APARCH parameters were obtained achieved through computationally expensive 3-dimensional grid searches. Autocorrelation functions for the ξ_t^2 series were calculated, for lags 1 – 50 hours, and the sum of their absolute values used as the metric to reflect the suitability of a parameter value set. Unfortunately investigations demonstrated that MLE parameter values were very drastically inaccurate. An alternative method was therefore tried – continuation of the grid-search, carefully exploring all local minima. The method was rendered successful, but expensive, after a more sophisticated success metric was adopted – still based on summed correlations.

Chapter 8:

Chapter 8 presented the development of a complete simulation methodology for the fitted model. In a reversal of order from the previous chapter, the method starts with suitable i.i.d. noise and adds layer upon layer of structure until wind speeds are finally produced. The development of the methodology was shaped by assessment of the simulation results at each stage of the process, and some additions to the model structure were deemed necessary. Many diverse aspects of the simulated series' dynamics were compared both with the historical series and with those generated from a simpler model. Finally, a realistic scenario was developed for the distribution of wind capacity in the fairly near future, so that simulated wind speeds could be converted to aggregated wind powers for this scenario.

8.1: A set of algorithms were developed and applied for generating an example 10—year synthetic series. The methodology's performance at each stage was assessed this lead to the development of an additional part of the model.

8.1.1: The first aspect of the methodology to be developed was an algorithm capable of generating temporally independent random deviates that have the same joint distribution in space as the unconditional residuals from the historic series. It was decided that a 20-dimensional, non-Gaussian copula is impossible to fit, and that a vine structure of pairwise non-Gaussian copulas, while possible to fit would represent an enormous challenge. It was therefore decided that at the outset that a Gaussian copula would be acceptable. A crude and mathematically dubious approach would was als tried that avoided dealing directly with the mathematics of 20-dimesnional copulas – and the approach, rather surprisingly, proved very successful.

8.1.2: An algorithm for generating conditionally heteroskedastic noise series with temporal structures given by the fitted APARCH model was developed and applied. This turned out to be much more challenging than simply applying the APARCH model structure to the unconditional noise series. Also, a simple algorithm was developed for re-establishing the inter- and intra-annual patterns in variance.

Initially an algorithm was developed that simply applied the APARCH model structure to randomly generated unconditional innovations. However it was found that at some early point in the series, following a set of large unconditional innovations, the variance exploded. This was true for each zone, and it was found that changing parameter values could only made the threshold for instability more extreme. Many statistical 'fudges' were tried, with none successful, therefore it became clear that the root of the problem was a built-in positive feedback, between the conditional variance and the conditional innovations. The fundamental nature of the algorithm had to be changed, involving the introduction of a new random variable. Getting this reformulation to work so took considerable effort, but a set of creative solutions were found.

A few extreme values of ξ_{t-1}, were produced that were much larger than any found in the historical series, and the decision was made that these had to be censored. Sound mathematical principles were used to decide that cut-off point, yet it felt like censorship might be betraying is some sense the original purpose of producing synthetic data synthetic data. A final re-scaling was necessary to match the synthetic innovations' long-term variance to that of the historical series, as was the case when generating the unconditional innovations. Simple multiplication did not work very well for achieving this re-scaling, but an innovative approach was developed. The final distributions of conditional series showed excellent agreement with historical ones. Inter-annual and intra-annual seasonality in variance were reintroduced.

Further re-scaling was necessary after that, but again distributions showed excellent agreement.

8.1.3: Algorithms were developed and applied for constructing wind speeds from the final synthetic noise. The first part of the algorithm filters noise through the VARMA(3,1) model, and as such was straightforward to construct.

The next stage was to reverse the 2nd differencing, to create the series $V_{i,t}$ from $U_{i,t}$, then reverse the 1st differencing to produce transformed wind speeds $X_{i,t}$. Some re-arranging of the equations defining the fractional differencing of a Gagenbauer process revealed that to generate a synthetic series for $X_{i,t}$, of length N, it is necessary to at $t = 1$, use synthetic values of $U_{i,t}$ for $1 \leq t \leq N$, and a segment of historical $V_{i,t}$ series for $-M_1 < t < 1$ to generate a synthetic series of $V_{i,t}$. Here, M_1 is a truncation limit large enough to cover several periods at the Gagenbauer frequency. Similarly, using this series along with a historical segment of $X_{i,t}$ generates the synthetic $X_{i,t}$ series. With both truncation limits set at 25,000 hours the simulation process was computationally very expensive, hence why only a single 10-year series was generated. Finally, deterministic seasonalities in mean were reintroduced – annual then diurnal.

8.1.4: After producing the synthetic wind speeds (but before the reversal of the Box-Cox transformation, it was necessary to develop and apply an algorithm to make final adjustments to the wind speed series.

The long-run mean of the synthetic series, at this stage of reconstruction, was found to be small but possibly 'unrealistic'. It must be noted however that this cannot be known definitively without historical series of annual mean wind speeds stretching back at least 100 years. Nonetheless, based on the available data, the mean of the synthetic series seems to be roughly 5 – 6 times greater than is realistic, apparently suggesting that fitted long memory parameters are too large. It was therefore decided that the deviation of the synthetic series at this point should be capped at 20% greater than the maximum deviation in the historical series, keeping the direction of deviation. Hopefully this is legitimate compensation for an inevitably imperfect fit, although it is also possibly another example of supressing the differences that make the generation of synthetic series preferable to the repeated sampling of historical series segments in a Monte Carlo simulation.

The synthetic series were all more leptokurtic than their historical counterparts. Various algorithms were developed to transform the synthetic distributions to closely resemble the historical ones, based on power transformations. A sophisticated algorithm was developed, after trying many possibilities. A simple algorithm was then developed for the final

reconstruction stages: adding long-term means, making all negative values zero, raising all values to the power 2.5 and rounding to the nearest knot.

8.2: An initial analysis of the synthetic series was conducted.

8.2.1: Marginal wind speed distributions were calculated for the synthetic wind speeds and compared to historical distributions.

It was found that mean values were in good agreement, medians were in excellent agreement, as were modes – except where the historical series mode is zero knots or close. Standard deviations were found to be in good agreement, the largest difference being 26%. Skewness was found to be the least similar of measures, with synthetic series significantly more skewed. Also the final synthetic series were more leptokurtic, despite the algorithm described above. However, when comparing distributions through visual inspections, the synthetic-historical pairs seem closely matched, suggesting that a few extreme values are to blame for the differences in summary statistics.

8.2.2: Plausible marginal and spatially aggregated wind power distributions were generated from the wind speeds and discussed, with the caveat that the conversion method was very crude. More than one up-scaling factor was used, to ensure generality. Marginal synthetic and historical series were found to be in very close agreement for all zones, particularly intermediate power levels. For the spatially aggregated series, the extent of agreement is lower than for individual zones, with the synthetic series under-representing the lowest and highest aggregate power states, i.e. < 10% and > 70%. This seems to imply an inadequacy of the assumed Gaussian copula. However their means and medians are in very close agreement, and their standard deviations are close.

The distributions of changes in the aggregate power were calculated for both series types, for 1 hour and 4 hour time differences. The results showed excellent agreement – and the fact that this is true may be considered one of the strongest validations of this research project's success.

8.2.3: A methodology was developed that allows examination of the distribution of multivariate states characterising in the wind speed field, allowing comparison of historical and synthetic joint distributions beyond simply the aggregated marginals. This methodology was rather complex but novel, and involved the definition of vector states for the wind speed field, and a comparison of their frequency of occurrence. The results were generally very good, although they do confirm a tendency for the synthetic series to under-represents states with a high spatial similarity, and over-represent highly mixed states.

8.3: A multivariate transition matrix model was fitted to the doubly-differenced series generated as part of the wind speed synthesizing process, at a subset of zones. The model was seen to consistently 'nudge' the VGARMA-APARCH model in such a way as to improve spatial joint distributions – although the impact of this addition was quite subtle. The very complex transition Markov chain model dealt with the previously identified problem of un-represented spatial correlations among innovations by working with transitions between vector states for the reduced wind speed field. These vector states also included the state of volatility at each location.

8.4: A thorough analysis was conducted on the dynamics of the synthetic and historical series, to further understand the extent of their similarity. A simpler VARMA model was also fitted for comparison, to gain some insight about establish whether the complexity of the 2-factor-VGARMA-APARCH is justified. This model kept the power transformation aspect, along with deterministic removal of seasonality in mean and variance on both diurnal and annual scales. However, long memory and conditional heteroskedasticity effects were ignored. The full model is referred to here as the Gagenbauer model, while the simpler one is referred to as the VAR model.

8.4.1: Autocorrelation functions were calculated and presented, with lag ranges of 0 – 500 hours (medium range) and 1001 – 10,000 hours (long range), for the historical series and those generated from the two models. For the medium range, the Gagenbauer series were found to be significantly better than the VAR series for most zones. For the long range, it was found that both synthetic model ACFs decayed more quickly than in the historic series, for all zones, and in many cases the two models were surprisingly similar. This suggests the opposite to most other comparisons: that the full model's long memory could be too light.

8.4.2: Several periodograms were prepared for discussion, for the historical series and those generated from the two models, focusing on different frequency ranges. The main observation made was that both series are in good agreement, but the Gagenbauer model is generally more successful.

Close to annual seasonality, for e.g. zone 1, the Gagenbauer match is indeed so good that the peaks cannot be distringuished (although VAR model is admittedly also good). For some zones, however, the VAR model seriously underestimates low frequency intensities. Between the annual and diurnal peaks, typically there is too much spectral intensity, for both models, but the Gagenbauer model is mostly worse. Both models are in excellent agreement about the height of the diurnal peaks.

8.4.3: Analysis was conducted on inter-annual variability and long-term means, for the historical series and those generated from the two models. The analysis considered: the variance of means, for January, July and annual means; long term means (for the same 3 periods); long term variances (for the 3 periods); and the variance of monthly variances (for the 3 periods). The results were mixed, but the Gagenbauer model outperformed the VAR model more often than not, overall. The results indicate that fitting two sets of long memory coefficients, one for the summer and another for winter, would have benefited the Gagenbauer model with regard to aspects tested here.

The correlations between monthly means and variances were compared, with both models doing well. The extent to which the synthetic series represent the wide range of monthly skewness values was examined. For several zones, their distributions were not recreated accurately by either model, although the Gagenbauer was found to be the best, on average.

8.5: A realistic wind capacity scenario for the fairly near future was developed. This was used to produce aggregated GB power outputs from the synthetic wind series. Assessments were made of how realistic the aggregated power output series are.

8.5.1: A methodology was developed and applied to represent the diurnal seasonality at hub height accurately. A simple approach was applied for speeding-up wind speeds from the recording station locations to realistic wind farm locations, also speeding-up from 10m to hub height. As previously acknowledged, the relationship between 10m and hub height wind speeds is much too complex to approach rigorously. However, the literature has established numerous precedents for the use of very simple approaches in this regard. Therefore the simplest possible approach was tested of merely replacing half of the diurnal seasonality in mean and variance during the simulation process. However, since this stage is followed by power transformation, the effect was to reduce the seasonality by a factor greater than a half. However, keeping the full seasonality in standard deviation and halving the mean proved reasonable and effective.

For location-to-location speed-up, adjusted valuesfrom the Bath wind model were adopted, specific to each location type within each zone. Bath wind model ratios were also used for speeding up to hub height.

8.5.2: A realistic scenario was constructed, for an unspecified time in the fairly near future, specifying the amount of wind generation capacity present at each type of location within each zone. This involved developing a policy on what type of current and planned future capacity may be excluded when constructing the scenario, to ease the classification process.

The scenario was based on the distribution of wind generation capacity currently operational, being constructed, having received planning consent and that have been submitted to the planning authorities. The scenario is for an unspecified time in the fairly near future when all wind-farms in the first three categories have been built, along with a random selection of half the capacity in the fourth category. The information was obtained by (time consuming) visual inspection of *Renewable UK*'s wind energy database. To ease burden, only projects > 25 MW considered (which does exclude a significant amount). Also, only offshore capacity that's close to shore was considered – specifically 12 nautical miles (22 km), i.e. projects within territorial waters. It was found that capacity was spread very unevenly between zones, ranging from 2,915 MW in zone 11 to only 2.6MW in zone 18. Offshore projects account for just over a third of capacity in this scenario, but would be considerably greater without the distance limit.

8.5.3: An algorithm was developed and applied for smoothing the single-location wind power outputs adequately, so that the aggregated output has realistic dynamics. Criteria were developed for testing whether this is the case. Wind speeds were transformed into power outputs using a 'softened' power curve, derived from the literature and the aggregated series analysed in relation to reported empirical observasions.

An algorithm for rendering more realistic the up-scaling of single location wind speeds to regional power outputs was found in the literature, and was adopted, with adjustments. The algorithm replaces hourly wind speeds with moving block averages – with 3 hour blocks found to be reasonable for GB.

The mean load factor for both versions of the series was 37.66%, and the median 35.57% and 35.55% for series 1 and 2 respectively. These are slightly higher than average observed values for GB, but given that 100% turbine availability was assumed, along with higher than historical proportion of offshore capacity, the results seem very positive. The series' s.d./mean ratios are, according to the literature, appropriate for a country the size of Denmark. This might be suggesting that more spatial smoothing is needed, although it could also be due to the concentration of most capacity in a small fraction of GB.

9.2. Future Work

The research conducted within this project must be understood as a leading, yet relatively small contribution to the process of developing the models and algorithms necessary for the sequential Monte Carlo simulation of entire power system. The need to engage with this much broader research challenge will become increasingly pressing as modern power systems move towards higher penetrations of renewable energy capacity, and as the ability to store energy and defer loads in large amounts become begins to be realised.

For this greater challenge, models would have to be fitted for each of the variables and algorithms constructed to generate realistic time series for them, along with reproducing their highly nonlinear interactions. The other crucial aspect for such simulation is the modelling of demand and its relationship to the meteorological variables. Advanced multivariate copula methods would surely be a large feature of such modelling work. That is a formidable modelling challenge, particularly with a large number of dimensions, but is one that the academic community must embrace. A methodology was proposed in chapter 6 that could act as a stepping-stone on the path towards such a full model.

A much smaller issue that could be addressed in future work is the extent to which model forecasting performance could be improved if other meteorological quantities were introduced as known exogenous variables – primarily wind direction, but also temperature, atmospheric pressure and perhaps circulation regime. Such variables would almost certainly not be used as linear regressors, rather they would affect model parameters for regression on past wind speeds. Rather than merely comparing simpler and more sophisticated models, the task would then be one comparing the value of model complexity of different type, one type being the use of many different variables. Another approach worth pursuing would be the simulation of wind vectors, i.e. simultaneously modelling and simulating wind speeds and directions – probably for a much smaller number of zones, due to the considerably greater complexity.

A significant problem exists with regard to validation of the model developed in this research. Its purpose is to provide more accurate values for quantities such as the overall reliability of a modelled power system, in terms of a metric such as LOLE, or e.g. the capacity credit of an additional GW of wind generation capacity. Until a full and accurate methodology for the full Monte Carlo simulation of a scenario is developed, then the accuracy of the wind speed modelling aspect cannot be fully known. Further, even if the methodology exists, with what value should the answer generated be compared, given that the original motivation is

that historical series are too short to be reliable? Perhaps insight could be gained by examining the behaviour of values obtained using historical wind speed series, with sections of the historical series sampled an increasing number of times to create a longer series. This could be contrasted with the behaviour of values obtained using synthetic series as the simulation length is increased. This would obviously require considerable computational resources.

Related to this is a recurring question that arose during the presentation of methodologies was whether the superiority of very long series is partially negated by the numerous censors and normalising transformations contained within the simulation algorithms. A definite answer to that question is unfortunately not possible without historical series of annual mean wind speeds stretching back at least 100 years.

A major alternative approach to the chosen long memory representation of low frequency variability, worthy of investigation, is a hidden Markov model. Such a model would probably require more than one layer of hidden structure – possibly one level representing the 'current' nature of atmospheric circulation over a wider area, in some sense, and another level determining the relative frequency of occurrence of those circulation types. Shortcomings of this approach are that a finite number of states cannot represent the smooth continuum of reality, and that in a multivariate context one would have to deal with vectors of hidden states. Relating these vector states to a single indicator variable, and understanding the dynamics of transitions between them, would be extremely challenging. Other investigations worth conducting would be to replace the VGARMA structure with one involving a combination of vector SARMA and ARFIMA model structures.

Several less fundamental differences in modelling approach are also worthy of investigation in future work. One is to follow the same procedures but: (i) without employing a power transformation; and (ii) taking logarithms instead. Another is to introduce seasonally varying long memory parameters – even two sets of values, one for summer and the other for winter, would probably be a considerable improvement.

Future work could potentially improve the modelling of conditional variance by fitting full APARCH models, i.e. with the power parameter free. The fitting of such models might be easier using the statistical software *R* than *Matlab*. It may be that with such models, particularly for values of the power parameter less than unity, the problem of exploding volatility during simulation would not occur, avoiding the need for a rather convoluted algorithm to deal with the problem.

With regard to the generation of regional wind power outputs from the single location wind speeds, there is plenty of scope for future work. One possibility is to explore simple but dynamic models for wind shear. Another is to model turbine availabilities in a more realistic way, i.e. offshore turbines experiencing a fault might remain offline until the end of a winter season. A useful line of enquiry would be how to realistically include further offshore projects within scenarios.

9.3. Summary Conclusions

The aim of this project was to build algorithms capable of generating synthetic wind speed fields across GB of arbitrary length, stochastic in nature but accurately reproducing spatio-temporal patterns on many different scales. The inherent complexity of working in a high-dimensional multivariate context places limitations on how accurate the individual series generated can be, but was necessary in order to obtain a reasonable spatial resolution. Whilst the results presented in chapter 8 are far from perfect, the fact that so many diverse aspects of the historical series are replicated reasonably well means that the modelling work may be considered successful.

The main self-appointed constraint for this project was that only past values of wind speeds could be used as explanatory variables; as such the project may be viewed as an exploration of the limits of such a model. Expressed in another way, the project examined the extent to which highly sophisticated model building can deliver synthetic series that are superior to those generated from simpler models. The answers found certainly defy easy characterisation. However, several rigorous observations about the best approach to the multivariate time series modelling of wind speeds may be made on the basis of this research. Arranged by modelling aspect, the observations are:

Seasonality: The seasonalities found in wind speed series are strong, complex and difficult to characterise, although they certainly comprise of both deterministic and stochastic components.

Heteroskedasticity: There is no doubt that the heteroskedastic nature of wind speeds should be reflected in the synthetic series. This includes annual and diurnal seasonalities in variance and random volatility clustering. Univariate models are entirely sufficient to model volatility clustering.

Long Memory: There is strong evidence to suggest that either long memory or hidden regime switching should be incorporated into a good time series model of wind speed. Due to difficulties associated with parameter fitting, it was seen that simpler models without long memory can be more accurate in reproducing some aspects of the historical series, although this does not constitute evidence that long memory is not truly present in the historical series. The 2-factor-Gagenbauer structure chosen here seems appropriate for a single multivariate model, and was adopted based on strong empirical evidence. However it is probably not the most accurate way of representing long memory for some zones. The alternative to the long memory representation of low frequency variability is a hidden Markov model, possibly with

more than one layer of hidden structure. The fitting of such a model would be extremely challenging.

Power Transformation: Since wind speed distributions are highly skewed, and often represented by the Weibull distribution family, several modelling options were possible with regard to initial transformation. In the multivariate case it is highly desirable that the series may be reasonably modelled as forming a MVN, which requires transformation.

Marginal and joint distribution of residuals: It was found that marginal error distributions, i.e. the distribution of model residuals/ noise innovations for individual zones, after accounting for conditional variance, deviate significantly from normality. They are significantly leptokurtic and as such it was found that modelling them as being Laplace or GED distributed is an improvement, depending on the zone. However a non-parametric characterisation was adopted here.

It was established that the joint distribution of residuals/ innovations may be characterised quite well by a Gaussian copula. However, situations where the majority of GB is very calm or very windy were found to be under-represented in favour of more mixed states. A model enhancement was developed that worked directly with vector states of zonal windiness and volatility, but was found to improve the situation only very subtly.

Conversion of speeds to zonal power outputs: It was established in the literature that conversion of single location wind speeds to regional/zonal power outputs is a highly complex process requiring sophisticated statistical methods and significant amounts of geographical information. As such doing so in a truly realistic manner was deemed beyond the scope of this project. However, fairly simple methodologies exist and provide reasonable approximations.

To summarise, this project has examined the nature of the GB wind resource in considerable detail, along with previous attempts to model it. A rigorous exploration was conducted on modelling options that could build upon and surpass existing work, with the final chosen solution being at the frontier of time series modelling. Results show that the model is successful in many regards but has several flaws, which is probably inevitable given the complex nature of meteorological systems.

Appendices

A1. Gagenbauer Frequencies and Differencing Parameters

Table A1.1 Gagenbauer Model Parameters

Zone	j_1	j_2	δ_1	δ_2
1	57	215	0.03	0.03
2	1	59	0.07	0.03
3	48	410	0.1	0.03
4	53	216	0.04	0.03
5	41	310	0.04	0.03
6	49	356	0.11	0.03
7	50	215	0.07	0.03
8	1	216	0.07	0.03
9	49	189	0.06	0.03
10	4	61	0.11	0.03
11	49	144	0.12	0.03
12	4	117	0.11	0.03
13	117	370	0.12	0.03
14	117	346	0.12	0.03
15	42	170	0.12	0.03
16	61	149	0.1	0.03
17	2	169	0.05	0.03
18	117	362	0.13	0.03
19	117	364	0.13	0.03
20	2	150	0.11	0.03

$j_1 =$ Fourier frequency, related to of the 1st Gagenbauer angular frequency through

$\omega_1 = 2\pi j_1/N.$

$j_2 =$ Fourier frequency, likewise related to of the 2nd Gagenbauer angular frequency.

$\delta_1 =$ fractional differencing parameter associated with the 1st Gagenbauer frequency.

$\delta_2 =$ fractional differencing parameter associated with the 2nd Gagenbauer frequency.

A2 Autoregressive and Moving Average Coefficient Matrices

Table A2.1 1st Autoregressive Coefficient Matrices

Columns 1 – 10.

1.03	0.09	0.12	-0.02	0.03	-0.03	0.04	0.01	0	0.07
0.02	0.85	-0.01	0.31	0	0.06	0	0.16	0	-0.09
-0.22	0.02	0.85	0.12	0.01	0.02	0.09	0.06	0.08	-0.01
-0.06	0.2	-0.1	1.01	-0.04	0.07	0.07	0.08	0.05	0.04
-0.05	0.29	0.09	0.08	0.97	0	-0.09	-0.07	0.05	-0.06
0	0.08	-0.16	0.52	-0.18	0.64	0.05	0.24	0.13	-0.06
-0.04	0.09	0.05	0.17	-0.09	0.07	0.86	0.09	0.07	0.09
0.04	0.1	0.02	0.39	-0.09	0	-0.05	1	0.05	0.14
-0.12	-0.12	0.04	0.19	-0.04	0.11	0.04	0.16	0.77	-0.07
-0.02	0.11	0.02	0.15	0.03	0.09	-0.04	0.29	0.06	0.51
0.07	0.12	-0.01	0.29	0.01	-0.13	0.02	0.02	-0.06	0.08
0.13	-0.28	-0.05	0.12	0.11	0.02	0	-0.01	0.2	0.18
0.21	0.05	-0.15	-0.04	0.13	0.04	-0.09	0.2	0.11	-0.13
-0.01	0.11	-0.01	0.07	-0.09	-0.11	-0.01	0	0.03	0.01
-0.04	-0.09	-0.02	-0.01	-0.04	0.09	0.18	-0.06	0.03	0.03
-0.09	-0.04	-0.02	-0.21	0	0.02	0.07	0.11	-0.06	-0.01
-0.02	0.05	0.02	0.2	0.05	-0.09	-0.04	0.02	0.06	-0.05
-0.01	-0.11	0.14	0.25	-0.1	-0.01	-0.29	0.09	0.06	0.05
0.19	-0.02	0.04	0.13	-0.09	-0.01	0.06	-0.13	0.01	0.08
-0.1	-0.04	-0.08	0.02	-0.05	-0.02	0.01	0.03	-0.04	-0.07

Columns 11 – 20.

-0.01	0.04	0.07	0.01	-0.06	0.01	0.06	0	0.07	-0.05
-0.04	-0.01	0.07	-0.02	-0.01	0.06	0.05	-0.05	0	-0.04
0	-0.01	0.11	-0.14	0.05	0.06	-0.03	0	0.11	-0.12
-0.03	-0.05	-0.03	-0.02	0.07	-0.05	0	0	0.01	0
0.1	0.03	0.09	-0.29	0.05	-0.07	0.05	0	0.06	-0.01
-0.02	-0.06	0.07	-0.09	0.05	0.07	0.04	-0.03	-0.06	0.03
-0.05	-0.03	-0.01	-0.02	0.12	-0.03	0.04	-0.05	0.01	0.04
0.12	-0.06	0.12	-0.08	-0.04	0.15	-0.03	0.12	-0.05	-0.01
0.19	-0.1	0.01	-0.03	-0.03	-0.09	0.03	-0.02	0.06	-0.04
0.19	0	0.04	0.06	0.03	-0.03	0.02	-0.09	0.07	0.11
0.95	-0.07	0.07	0.07	0.08	-0.19	-0.02	0.01	0.03	-0.05
-0.02	0.5	0.13	-0.04	-0.05	0.01	-0.12	-0.02	0.06	0.24
0.04	-0.06	0.77	-0.03	0.15	0.27	0.12	-0.06	0.06	0.25
0.13	0	0	0.9	0.02	-0.02	0.01	0.01	0	-0.03
0.03	-0.02	0.12	0.31	0.68	-0.09	0.01	0.07	0.06	-0.1
0.04	0.09	0.06	0.1	-0.01	0.77	0.03	0.03	0.1	0.13
-0.06	-0.01	0.03	0.11	-0.05	0	0.79	0.09	0	0.02
0.01	-0.07	0	-0.03	0.26	0.29	0.16	0.88	0	0.01
0.11	-0.03	0.12	-0.2	0.04	-0.03	0.27	0.1	0.67	0.11
-0.02	0	0.01	0.01	0.1	-0.06	0.1	0.1	0.15	0.86

Table A2.2 2nd Autoregressive Coefficient Matrix

Column 1 – 10.

-0.17	-0.04	-0.03	0.02	-0.02	0.01	-0.03	-0.01	0.01	-0.03
-0.01	-0.06	0.01	-0.17	0	-0.03	-0.02	-0.09	0	0.04
0.21	0.02	-0.08	-0.1	0.01	-0.01	-0.06	-0.03	-0.05	0
0.06	-0.09	0.05	-0.17	0.04	-0.03	-0.03	-0.05	-0.02	-0.02
0.05	-0.18	0	-0.07	-0.17	0.01	0.07	0.06	-0.02	0.03
0.01	-0.05	0.1	-0.33	0.13	-0.06	-0.01	-0.12	-0.08	0.02
0.03	-0.07	-0.03	-0.08	0.05	-0.02	-0.13	-0.04	-0.03	-0.05
-0.02	-0.06	-0.01	-0.28	0.06	0	0.04	-0.21	-0.03	-0.06
0.1	0.06	-0.01	-0.1	0.03	-0.04	0.05	-0.08	-0.12	0.05
0.01	-0.1	-0.02	-0.12	-0.02	-0.03	0.02	-0.1	0	0
-0.06	-0.09	0.01	-0.22	0	0.05	-0.01	-0.01	0.05	-0.03
-0.08	0.18	0.03	-0.09	-0.05	-0.01	-0.02	0.03	-0.12	-0.04
-0.17	-0.02	0.06	0.01	-0.08	-0.02	0.04	-0.13	-0.06	0.07
0.01	-0.08	0.01	-0.05	0.06	0.05	0.02	-0.01	-0.01	0
0.04	0.07	0.01	-0.02	0.03	-0.05	-0.12	0.05	-0.01	0
0.07	0.04	0.02	0.16	0	0	-0.03	-0.07	0.03	0
0.02	-0.04	-0.01	-0.15	-0.04	0.03	0.02	-0.03	-0.03	0.03
0.02	0.06	-0.07	-0.16	0.06	0.01	0.15	-0.07	-0.04	-0.02
-0.16	0.01	-0.02	-0.12	0.05	0.01	-0.03	0.05	-0.01	-0.02
0.08	0.03	0.05	0	0.05	0	0	-0.02	0.02	0.03

Columns 11 – 20.

0	-0.02	-0.04	-0.01	0.01	-0.01	-0.04	0.01	-0.04	0.02
0.03	0.01	-0.04	0.01	0	-0.03	-0.04	0.03	-0.01	0.01
-0.02	0	-0.06	0.07	-0.03	-0.03	0.03	0	-0.05	0.05
0.01	0.02	0.01	0.01	-0.03	0.02	0.01	0	-0.01	0
-0.04	-0.01	-0.05	0.16	-0.02	0.02	-0.03	-0.01	-0.03	-0.01
0	0.02	-0.04	0.03	-0.02	-0.04	-0.03	0.01	0.02	-0.03
0.02	0.01	0	0.02	-0.04	0.02	-0.02	0.03	-0.01	-0.02
-0.06	0.03	-0.07	0.03	0	-0.08	0.01	-0.05	0.01	-0.02
-0.06	0.06	0	0.04	0.02	0.05	-0.01	0.01	-0.02	0.02
-0.09	0.02	-0.02	-0.05	-0.01	0.02	-0.03	0.05	-0.02	-0.07
-0.21	0.04	-0.02	-0.01	-0.03	0.1	0.03	0	-0.01	0.04
0.05	-0.01	-0.03	0.05	0.05	0	0.1	0.01	-0.04	-0.16
-0.01	0.06	-0.14	0.02	-0.02	-0.12	-0.08	0.01	-0.04	-0.17
-0.04	0	0	-0.18	-0.01	0.02	0.03	0	0	0.04
-0.01	0.02	-0.05	-0.16	-0.09	0.05	-0.01	-0.02	0	0.05
-0.03	-0.03	0	-0.06	0.02	-0.11	-0.01	-0.01	-0.03	-0.07
0.05	0	-0.01	-0.02	0.01	0.01	-0.03	-0.04	0	-0.01
-0.03	0.03	-0.03	0.02	-0.09	-0.14	-0.1	-0.2	0.01	-0.01
-0.05	0.02	-0.03	0.11	0.03	0.03	-0.16	0	-0.1	-0.05
0	0	0.01	0	-0.05	0.07	-0.07	-0.03	-0.05	-0.16

Table A2.3 3rd Autoregressive Coefficient Matrix

Columns 1 – 10.

10^{-3} x

11	3	2	-12	4	-2	4	-9	-12	0
-2	15	-7	-26	1	-4	-11	-2	-5	-2
-25	27	26	7	-15	-5	0	-11	-9	-6
-7	-11	8	-1	5	-9	-3	-19	-11	-1
-13	-7	-4	21	13	-11	4	0	12	-13
-9	6	31	-30	-1	26	-2	-4	-2	4
4	-5	5	-12	6	-6	13	-15	-3	0
-8	-14	-3	-25	8	-5	-5	10	-6	-7
-4	17	-13	20	4	-5	-20	-10	44	2
-4	-9	1	-16	-4	4	-16	13	13	29
-10	-7	-9	5	3	7	-2	11	-5	-2
-9	30	-23	22	3	-10	-19	-8	10	-5
-7	-24	1	8	-10	-5	1	21	-5	-1
2	-3	4	1	11	4	4	1	-3	-3
-1	6	6	-2	8	-12	-8	-6	6	-3
8	9	1	-2	-4	-3	3	-11	3	2
-13	-13	-9	-6	-1	0	8	19	6	-4
-7	6	-8	-15	4	6	1	11	-1	-9
-6	-9	-4	10	10	3	-2	8	-1	-17
5	9	17	8	-2	8	6	1	-4	3

Columns 11 – 20.

10^{-3} x

4	2	1	-4	0	-10	-10	-6	1	3
0	-6	-2	-3	2	4	-1	0	1	11
9	6	2	4	-16	-9	-3	1	3	3
4	0	6	3	1	3	10	-2	3	0
2	-5	10	2	-11	-4	12	4	4	16
-1	5	14	11	3	2	-14	7	2	4
-2	-4	-1	0	-3	-4	12	-1	0	0
4	-9	1	-6	7	-15	0	-1	-4	11
4	-7	2	11	13	-6	-1	-7	-1	11
-7	-4	9	-2	16	-4	-8	-8	-5	1
35	1	12	-7	1	-3	-19	6	7	-2
20	35	4	14	5	-22	-19	-10	-3	0
-6	-5	15	-6	8	-23	-6	-12	-18	1
7	-1	5	14	0	-2	6	6	5	-16
9	1	4	20	2	10	-6	3	-4	-1
11	0	1	-1	-2	5	-4	-2	1	10
8	6	1	1	3	-3	40	5	-5	-6
-5	-6	0	-12	-5	-2	25	27	-2	-15
-5	-10	-2	-5	0	0	-3	12	30	9
2	4	3	-15	0	-7	2	16	-2	25

Table A2.4 Moving Average Coefficient Matrix

Columns 1 – 10.

-0.2	-0.08	-0.09	0.02	-0.01	0.03	-0.03	0	-0.01	-0.07
0	-0.18	0.03	-0.19	0.01	-0.06	0.01	-0.13	0	0.09
0.26	0.03	-0.34	-0.1	0.04	-0.02	-0.07	-0.05	-0.09	0.01
0.06	-0.13	0.1	-0.21	0.05	-0.06	-0.02	-0.05	-0.03	-0.03
0.07	-0.27	-0.02	-0.06	-0.31	0	0.1	0.1	-0.04	0.05
-0.01	-0.05	0.17	-0.39	0.17	-0.2	0.04	-0.2	-0.12	0.08
0.04	-0.08	-0.05	-0.07	0.09	-0.04	-0.2	-0.06	-0.02	-0.09
-0.03	-0.09	-0.01	-0.35	0.1	0.02	0.09	-0.33	-0.02	-0.12
0.11	0.13	-0.04	-0.12	0.05	-0.09	0.09	-0.13	-0.15	0.1
0.02	-0.12	-0.04	-0.14	-0.01	-0.07	0.05	-0.18	-0.01	-0.05
-0.08	-0.11	0.01	-0.24	0	0.11	-0.02	0.01	0.1	-0.08
-0.13	0.28	0.06	-0.12	-0.08	-0.01	0	0.02	-0.19	-0.14
-0.2	-0.04	0.15	0.03	-0.11	-0.03	0.09	-0.19	-0.11	0.13
0.02	-0.12	0	-0.04	0.09	0.11	0.03	-0.01	-0.02	-0.01
0.04	0.08	0.03	0	0.04	-0.08	-0.17	0.06	-0.01	-0.02
0.08	0.04	0.02	0.21	0	-0.01	-0.06	-0.1	0.06	0.01
0.02	-0.06	-0.02	-0.19	-0.05	0.09	0.06	-0.02	-0.05	0.04
0	0.13	-0.14	-0.25	0.09	0.01	0.27	-0.08	-0.06	-0.04
-0.19	0.03	-0.05	-0.14	0.09	0.02	-0.06	0.12	-0.02	-0.07
0.1	0.04	0.07	-0.01	0.06	0.02	0	-0.05	0.04	0.07

Columns 11 – 20.

0.01	-0.04	-0.07	-0.02	0.07	0	-0.06	0	-0.07	0.05
0.05	0.01	-0.06	0.02	0.01	-0.06	-0.05	0.05	-0.01	0.03
0	0.02	-0.1	0.15	-0.05	-0.05	0.03	-0.01	-0.11	0.11
0.04	0.05	0.03	0.04	-0.07	0.05	0.02	-0.01	-0.01	0
-0.09	-0.03	-0.09	0.29	-0.05	0.08	-0.04	0	-0.05	0.02
0.04	0.06	-0.07	0.1	-0.05	-0.07	-0.05	0.03	0.05	-0.03
0.06	0.04	0.02	0.05	-0.11	0.02	-0.03	0.05	-0.01	-0.03
-0.09	0.06	-0.12	0.08	0.05	-0.14	0.03	-0.12	0.04	0.01
-0.14	0.11	-0.01	0.07	0.05	0.09	-0.02	0.01	-0.06	0.04
-0.15	0.02	-0.04	-0.04	0	0.05	-0.02	0.09	-0.07	-0.11
-0.35	0.09	-0.06	0	-0.07	0.2	0.04	-0.01	-0.04	0.04
0.05	-0.02	-0.1	0.04	0.11	0.01	0.13	0.01	-0.05	-0.23
-0.02	0.09	-0.29	0.04	-0.1	-0.17	-0.11	0.06	-0.04	-0.21
-0.07	0	0	-0.25	0	0.02	0.05	0	0	0.03
0.01	0.04	-0.1	-0.24	-0.24	0.13	-0.01	-0.06	-0.03	0.11
-0.03	-0.08	0	-0.09	0.03	-0.22	-0.02	-0.01	-0.06	-0.08
0.08	0	-0.03	-0.05	0.06	0.01	-0.07	-0.08	0	-0.01
-0.01	0.07	0.01	0.06	-0.24	-0.27	-0.11	-0.39	0.05	0.04
-0.12	0.03	-0.1	0.2	0.02	0.09	-0.24	-0.02	-0.2	-0.04
0.01	0	0	-0.01	-0.09	0.12	-0.09	-0.07	-0.1	-0.25

A3 AGARCH Model Parameters

Table A3.1 APARCH Model Parameters

Zone	α_0	α_1	λ	β_0
1	0.0126	0.28	0.84	0.41
2	0.0140	0.38	0.71	0.41
3	0.0078	0.31	0.95	0.36
4	0.0134	0.19	0.37	0.63
5	0.0150	0.18	0.68	0.53
6	0.0080	0.09	0.61	0.74
7	0.0090	0.09	0.94	0.65
8	0.0100	0.10	0.75	0.61
9	0.0025	0.18	0.94	0.65
10	0.0063	0.20	0.70	0.65
11	0.0190	0.24	0.33	0.65
12	0.0105	0.12	0.87	0.68
13	0.0170	0.21	0.43	0.67
14	0.0007	0.41	0.72	0.44
15	0.0100	0.09	0.55	0.65
16	0.0161	0.11	0.79	0.64
17	0.0057	0.29	0.84	0.50
18	0.0020	0.25	0.83	0.62
19	0.0088	0.18	0.64	0.66
20	0.0168	0.11	0.65	0.65

A4 Final Wind Speed Distributions

Table A4.1 Final Wind Speed Distribution Percentiles

Series	10%	20%	30%	40%	50%
Historical Zone 2	3	5	7	9	11
Historical Zone 7	4	6	8	9	11
Historical Zone 15	2	3	5	6	7
Historical Aggregated	111	135	156	176	198
Synth. Zone 2	4	6	8	10	11
Synth. Zone 7	4	6	8	9	11
Synth. Zone 15	2	4	5	6	7
Synth. Aggregated	113	142	166	188	210
Synth. Trans. Zone 2	4	6	8	10	11
Synth. Trans. Zone 7	4	6	8	9	11
Synth. Trans. Zone 15	2	4	5	6	7
Synth. Trans. Aggregated	113	143	166	188	209

Continued…

Series	60%	70%	80%	90%
Historical Zone 2	12	14	17	21
Historical Zone 7	13	15	18	22
Historical Zone 15	9	10	12	15
Historical Aggregated	221	247	281	332
Synth. Zone 2	13	16	18	22
Synth. Zone 7	13	16	19	23
Synth. Zone 15	9	10	12	16
Synth. Aggregated	233	259	291	340
Synth. Trans. Zone 2	13	15	18	22
Synth. Trans. Zone 7	13	16	18	23
Synth. Trans. Zone 15	9	10	12	15
Synth. Trans. Aggregated	232	258	291	341

Synth. = Synthetic series, Synth. Trans. = Synthetic series enhanced by transition matrix model

Table A4.2 Final Wind Speed Distribution Moments

Series	Mean	Skewness	Kurtosis
Historical Zone 2	11.51	0.87	3.88
Historical Zone 7	12.10	0.66	3.32
Historical Zone 15	8.11	0.81	3.87
Historical Aggregated	211.47	0.84	3.68
Synth. Zone 2	12.42	0.88	4.16
Synth. Zone 7	12.60	1.00	4.60
Synth. Zone 15	8.26	0.94	3.87
Synth. Aggregated	220.35	0.81	4.42
Synth. Trans. Zone 2	12.40	1.17	5.87
Synth. Trans. Zone 7	12.62	1.21	5.66
Synth. Trans. Zone 15	8.25	1.11	4.95
Synth. Trans. Aggregated	220.56	0.96	5.11

A5 Example Matlab Scripts

A5.1 Part of the Simulation Algorithm – Fractional Integration

```
% Program to reverse fractional differencing

load('frac_diff.mat');
load('d1_list.mat');
load('v1_list.mat');
load('v2_list.mat');
load('declimatised_series.mat');
load('max_dev.mat');

ls = 87672;
d2 = -0.03;
first_integ = zeros(20,ls+25000);
first_integ(:,1:25000) = frac_diff(1:20,116550:141549);

for j = 1:20
   v = cos(2*pi*v2_list(j)/175297);
   d = d2;
   gagen  = zeros(1,25001);
   gagen(1) = 1;
   gagen(2) = 2*d*v;

   for i = 2:25000
      gagen(i+1) = 2*v*(((d-1)/i) + 1)*gagen(i) - ((2*(d-1)/i)+1)*gagen(i-1);
   end
   gagen(2:25001) = -1*gagen(2:25001);

   for i = 25001:(ls+25000)
      list = (i-1):-1:(i-25000);
      first_integ(j,i) = synth_full_frac_trans(j,i-25000) + first_integ(j,list) * gagen(2:25001)';
   end
   save first_integ.mat first_integ
end

first_integ = first_integ(:,25001:(ls+25000));
save first_integ.mat first_integ

full_integ = zeros(20,ls+25000);
full_integ(:,1:25000) = declimatised_series(1:20,141550:166549);

for j = 1:20
   v = cos(2*pi*v1_list(j)/175297);
   d = -d1_list(j);
   gagen  = zeros(1,25001);
   gagen(1) = 1;
   gagen(2) = 2*d*v;
```

```
    for i = 2:25000
        gagen(i+1) = 2*v*(((d-1)/i) + 1)*gagen(i) - ((2*(d-1)/i)+1)*gagen(i-1);
    end
    gagen(2:25001) = -1*gagen(2:25001);

    for i = 25001:(ls+25000)
        list = (i-1):-1:(i-25000);
        full_integ(j,i) = first_integ(j,i-25000) + full_integ(j,list) * gagen(2:25001)';
    end

    full_integ(j,:) = (mean(abs(declimatised_series(j,:)))/mean(abs(full_integ(j,:))))*full_integ(j,:);

    if mean(full_integ(j,:)) > max_dev(j)
        full_integ(j,:) = full_integ(j,:) - mean(full_integ(j,:)) + max_dev(j);
    elseif mean(full_integ(j,:)) < -1*max_dev
        full_integ(j,:) = full_integ(j,:) - mean(full_integ(j,:)) - max_dev(j);
    end

    save full_integ.mat full_integ
end
full_integ_trans = full_integ(:,25001:(ls+25000));
full_integ_trans = [full_integ_trans; synth_full_frac_trans(21:24,:)];

save full_integ_trans.mat full_integ_trans
```

A5.2 Conversion of Simulated Wind Speeds to Powers in Example Scenario

```
% Program to establish the power curve for winds in knots and convert wind speeds
% into power outputs for the example scenario.


power_curve_s1 = zeros(1,97);
power_curve_s = zeros(1,200);

ls = 87672;
coastal_powers = zeros(20,ls);
lowland_powers = zeros(20,ls);
upland_powers = zeros(20,ls);
offshore_powers = zeros(20,ls);
balanced_powers = zeros(20,ls);
final_aggregate_p = zeros(1,ls);

ms_list = 0.5144*(0:57);

power_curve_s1(3) = 0.0025;
power_curve_s1(4) = 0.0100;
power_curve_s1(5) = 0.0250;
power_curve_s1(6) = 0.0750;
power_curve_s1(7) = 0.1335;
power_curve_s1(8) = 0.2000;
power_curve_s1(9) = 0.3500;
power_curve_s1(10) = 0.4200;
power_curve_s1(11) = 0.6000;
power_curve_s1(12) = 0.7900;
power_curve_s1(13) = 0.8650;
power_curve_s1(14) = 0.9500;
power_curve_s1(15) = 0.9700;
power_curve_s1(16) = 0.9900;
power_curve_s1(17) = 0.9950;
power_curve_s1(18) = 0.9975;
power_curve_s1(19:22) = ones(1,4);
power_curve_s1(23) = 0.9950;
power_curve_s1(24) = 0.9850;
power_curve_s1(25) = 0.9150;
power_curve_s1(26) = 0.6900;
power_curve_s1(27) = 0.3600;
power_curve_s1(28) = 0.1000;
power_curve_s1(29) = 0.0200;


for j = 1:58
   power_curve_s(j) = ((ceil(ms_list(j)) - ms_list(j))*power_curve_s1(floor(ms_list(j))+1)) +
((ms_list(j) - floor(ms_list(j)))*power_curve_s1(ceil(ms_list(j))+1));
end
```

```
 for i = 1:20
    coastal_powers(i,:) = power_curve_s(coastal_speeds(i,:)+1);
    lowland_powers(i,:) = power_curve_s(lowland_speeds(i,:)+1);
    upland_powers(i,:) = power_curve_s(upland_speeds(i,:)+1);
    offshore_powers(i,:) = power_curve_s(upland_speeds(i,:)+1);

    balanced_powers(i,:) = capacities(i,1)*coastal_powers(i,:) +
capacities(i,2)*lowland_powers(i,:);
    balanced_powers(i,:) = balanced_powers(i,:) + capacities(i,3)*upland_powers(i,:) +
capacities(i,4)*offshore_powers(i,:);

    final_aggregate_p = final_aggregate_p + balanced_powers(i,:);
 end
```